https://pubsonline.informs.org/journal/deca

DECISION ANALYSIS
Articles in Advance, pp. 1–19
ISSN 1545-8490 (print), ISSN 1545-8504 (online)

Harnessing Human Uncertainty to Train More Accurate and Aligned AI Systems

Gunnar P. Epping,[a,b,c,*] Andrew Caplin,[d] Erik Duhaime,[c] William R. Holmes,[a,e] Daniel Martin,[f] Jennifer S. Trueblood[a,b]

[a] Cognitive Science Program, Indiana University, Bloomington, Indiana 47405; [b] Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana 47405; [c] Centaur.ai, Boston, Massachusetts 02109; [d] Department of Economics, New York University, New York, New York 10012; [e] Department of Mathematics, Indiana University, Bloomington, Indiana 47405; [f] Department of Economics, University of California, Santa Barbara, California 93106
*Corresponding author

Contact: gunnarepping@gmail.com, https://orcid.org/0000-0003-2445-2573 (GPE); ac1@nyu.edu (AC); erik@centaurlabs.com (ED); wrholmes@iu.edu (WRH); daniel@martinonline.org, https://orcid.org/0000-0001-6483-3923 (DM); jstruebl@iu.edu (JST)

Received: May 14, 2025
Revised: September 28, 2025
Accepted: October 23, 2025
Published Online in Articles in Advance: April 1, 2026

https://doi.org/10.1287/deca.2025.0395

Abstract. Artificial intelligence (AI)-augmented decision making (AIADM) aims to leverage the computational power of machine learning (ML) models to assist humans in their decision-making processes. In many such systems, especially for complex tasks like medical image classification, ML models are often trained on large data sets annotated by humans. Neglecting to account for human decision-making biases when constructing these labeled data sets can lead to biased data sets, and subsequently models trained on such data sets can inherit the biases. We propose a novel approach to developing AIADM systems that aims to overcome these challenges by harnessing human uncertainty. Our approach has three elements: We collect subjective judgments from human annotators, we calibrate those subjective judgments, and we use the recalibrated subjective judgments to create probabilistic (i.e., soft) labels, which the AI decision aid is then trained on. We evaluate our methods through two studies using data from DiagnosUs, a crowdsourcing platform for medical image annotation. Across multiple training data sets, we assess how our proposed methods impact three key properties of AI decision aids that could benefit from leveraging human uncertainty in data annotation: accuracy, calibration, and alignment with human uncertainty. Our results show that ML models trained on recalibrated soft labels are more accurate and better aligned with expert judgments. We also observe a tradeoff between ML calibration and alignment with human uncertainty. These findings highlight the value of capturing and correcting human uncertainty in ML training data when developing AI systems.

History: This paper has been accepted for the *Decision Analysis* Special Issue on the Implications of Advances in Artificial Intelligence for Decision Analysis.
Funding: This work was supported by the Alfred P. Sloan Foundation (Cognitive Economics at Work).
Supplemental Material: The online appendix is available at https://doi.org/10.1287/deca.2025.0395.

Keywords: calibration • wisdom of the crowds • biases • artificial intelligence • subjective probability

1. Introduction

Artificial intelligence (AI)-augmented decision-making (AIADM) describes a situation where an AI decision aid provides information to a human decision maker to assist them in their decision-making process (Murray et al. 2021). For example, in the context of medical diagnostics, an AI decision aid can be used to estimate the likelihood of possible diagnoses, and the human decision maker can use the estimates to help arrive at a final diagnosis (Lee et al. 2020).

AIADM can lead to more accurate decisions compared with either human or AI in isolation (Cacciamani et al. 2023, Lu and Zhang 2024, Park et al. 2024). There are two main reasons why AIADM improves decision making. First, the decision aid can help improve choices when humans and machines have uncorrelated learning from the same information (Steyvers et al. 2022). Second, the decision aid can help reduce the effort or time that a human has to put into a problem, which can free them up to look at other information

that is not available to the machine (Grimon and Mills 2025).

However, AIADM can backfire if the AI decision aid is incorrectly certain. When this happens, decision quality erodes because humans may follow it when they should not. Behavioral biases can exacerbate these concerns, such when decision-makers are underconfident (Caplin et al. 2024) or when they overly follow AI (Agarwal et al. 2023). On top of this, incorrect certainty can cause humans to stop using an AI decision aid entirely. For example, in a phenomenon known as algorithm aversion (Dietvorst et al. 2015), people have been shown to lose trust in models after seeing them err. Therefore, both the accuracy and calibration (which relates to the certainty expressed by AI models) of AI decision aids influence how human interact with them.

Aside from the accuracy of the decision aid, Grgić-Hlača et al. (2022) demonstrated that people are more willing to revise their decisions, leading to a greater improvement in accuracy, when an AI decision aid produces judgments similar to their own. Therefore, improving the alignment between human judgments and the output of AI decision aids can increase people's willingness to integrate the information provided by decision aids into their decision-making process. Because the improved accuracy of AIADM depends on humans selectively integrating information from AI decision aids, enhancing how receptive human decision makers are to these aids is a critical component of making AIADM effective.

Based on this previous work, three key properties that influence how humans interact with AI decision aids are model accuracy, calibration, and alignment with expert judgments. Unfortunately, AIADM systems are often developed without considering all three of these properties or how they may trade off. The development of an AIADM system often involves multiple stages, which may include collecting data annotations from human raters, aggregating those annotations to create a labeled data set, and designing and training a machine learning (ML) model to serve as the decision aid. Therefore, people are not only central to using AIADM systems but also deeply involved in their design and training. Neglecting to account for human decision-making biases when constructing labeled data sets can lead to biased data sets, and subsequently models trained on such data sets can inherit these biases (Bolukbasi et al. 2016, Bender et al. 2021). These models can perform suboptimally as their biases lower their accuracy and reliability when deployed in an AIADM system.

We propose a novel approach to developing AIADM systems that better accounts for these properties. This approach departs from standard practice in three ways. First, we collect subjective judgments from human annotators. Second, we recalibrate those subjective judgments. Third, we use the recalibrated subjective judgments to create probabilistic (i.e., soft) labels, which the AI decision aid is then trained on.

Critically, our approach is based on an understanding of the decision-making process of human annotators whose judgments are used to generate labeled data sets. These humans raters operate in the context of crowdsourcing, which is one of the most popular methods for acquiring labeled data sets (Deng et al. 2009). In a phenomenon known as the "Wisdom of the Crowd" (Galton 1907, Surowiecki 2005), crowdsourcing leverages the knowledge of a diverse group of individuals to produce group decisions that are more accurate on average compared with the decisions of a randomly selected individual within the group. The Wisdom of the Crowd can be so effective that crowdsourced data sets from novice annotators can outperform those created by a single expert, even in domains like medical image diagnostics, where expertise was previously assumed essential (Maier-Hein et al. 2014, Duhaime et al. 2023, Hasan et al. 2024b).

When crowdsourcing labels for classification tasks, individual annotators are typically asked to provide categorical responses that reflect what they believe to be the most likely classification (Collins et al. 2022). According to the majority voting algorithm (Sorkin et al. 2001, Kurvers et al. 2016), which weighs the decisions of each annotator equally, the consensus label for a given data instance is the classification with the most votes across annotators. The majority voting algorithm has been shown to be very robust and effective despite its simplicity (Hastie and Kameda 2005, Kameda et al. 2011). However, using categorical response options may not be the best elicitation method, because categorical responses do not record people's confidence in their judgments (Winkler et al. 2019). Having some measure of confidence is valuable because people's confidence is

typically correlated with the likelihood of their judgment being correct (Murphy 1972, Baranski and Petrusic 1998, Pleskac and Busemeyer 2010, Vickers 2014), a finding referred to as the resolution of confidence. Hence, having a measure of confidence to accompany classifications provides information that can be leveraged to improve the accuracy of crowdsourced labels when aggregating annotations across individuals. For example, in confidence-weighted majority voting, each individual's vote is weighted by how confident they are in their classification, which can lead to more accurate group decisions than simple majority voting (Nitzan and Paroush 1982, Grofman et al. 1983, Meyen et al. 2021). Alternatively, the maximum confidence slating algorithm defines the consensus classification as the classification of the individual with the highest confidence (Koriat 2012), which can also improve consensus accuracy over majority voting. Therefore, having some measure of confidence to accompany annotators' classifications can improve the accuracy of crowdsourced data sets.

Unlike categorical responses, subjective probability judgments allow people to simultaneously encode both their classification and their confidence in that classification. This approach works particularly well for tasks with binary outcomes. For example, if people are asked to estimate the likelihood that a medical image contains a cancerous white blood cell, their classification (whether the image contains a cancerous cell) is based on whether their subjective probability judgment is above 50%, and their confidence is based on the distance between their subjective probability judgment and 50%. Because confidence-weighted wisdom of the crowd algorithms weigh high-confidence judgments more heavily compared with low-confidence ones, the improved accuracy afforded by these algorithms compared with that of the simple majority voting algorithm depends upon the calibration of the judgments. Subjective probability judgments are considered well calibrated if the judged probability of an event matches the relative frequency of the event (Koehler et al. 2002).

However, it is well documented that human probability judgments are often poorly calibrated (Lichtenstein et al. 1977, Griffin and Tversky 1992). Four human decision-making biases that degrade the calibration of subjective probability judgments are overprediction, underprediction, overextremity, and underextremity biases (Griffin and Brenner 2004). An overprediction bias occurs when people systematically overestimate the likelihood of an event, whereas an underprediction bias occurs when people systematically underestimate the likelihood of an event. An overextremity bias occurs when people overuse the extreme ends of the scale (judgments below 50% are too low and judgments above 50% are too high), whereas an underextremity bias occurs when people's judgments are too conservative (judgments below 50% are too high and judgments above 50% are too low). Individuals can exhibit one of these biases or a combination of over/underprediction and over/underextremity biases.

The second feature of our approach is to minimize the presence of these biases in subjective probability judgments by recalibrating them via the linear in log odds (LLO) function (Tversky and Fox 1995, Birnbaum and McIntosh 1996, Turner et al. 2014), which transforms subjective probability judgments into probability judgments that are more aligned with objective truth. Recalibration requires a holdout set of judgments on images for which we have ground truth labels. Recalibration works by fitting the LLO function to the holdout set of judgments, and then using the fitted function to transform the entire set of judgments. Importantly, the LLO function has two free parameters: one to counteract over/underprediction biases and another to counteract over/underextremity biases (Gonzalez and Wu 1999).

The third feature of our approach is to improve the alignment between ML models and human judgments by training ML models on probabilistic (known as "soft") labels rather than deterministic (known as "hard") labels. Hard label data sets contain strictly categorical labels for each data instance, whereas soft label data sets contain probability distributions over all possible labels for each data instance (Nguyen et al. 2014). For example, if an AI decision aid is trained to predict whether an image contains a cancerous white blood cell, and 60% of annotators label the image as "cancerous," the soft training label would reflect this distribution: 60% "cancerous" and 40% "not cancerous." In contrast, the hard label would simply be "cancerous." Unlike hard labels, soft labels all one to encode uncertainty in the classification, which makes them especially advantageous in fields such as medical diagnostics where ground truth can be difficult or impossible to obtain.

An additional benefit of crowdsourcing is that soft labels obtained through this process reflect human uncertainty in the classification (Peterson et al. 2019). Training models to output judgments that reflect this uncertainty can lead to greater alignment between AI decision aids and human judgments. Nevertheless, given that crowdsourced labels are probability distributions, they can exhibit the same biases that plague subjective probability judgments elicited from individual annotators. Even if the judgments from individual annotators are perfectly calibrated, the crowd labels created by aggregating these judgments can exhibit over/underextremity biases (Hora 2004, Ranjan and Gneiting 2010), so recalibrating individuals' subjective probability judgments does not circumvent this issue. Similar to how we recalibrate individual judgments, we also recalibrate crowdsourced labels using the LLO model to ensure that the soft labels are not only reflective of human uncertainty but are also well-calibrated.

To recap, we propose methods that explicitly consider accuracy, calibration, and alignment in the development and training of an AI decision aid. We will evaluate the efficacy of our proposed methods using data from two studies. In both studies, people evaluate whether an image of a white blood cell contains a cancerous white blood cell (a blast cell). We chose this task for several reasons. First, given that AIADM systems are deployed in real-world settings, we wanted to study AIADM in the context of a task that reflects one of these settings. Second, novices can achieve above chance accuracy in this task with minimal training, which is essential to the success of any crowdsourcing task, since crowdsourcing workers often have limit domain expertise. Third, the task is still difficult enough that experts make mistakes and express uncertainty in their classifications. In Study 1, we analyze binary classifications and subjective probability judgments, obtained via DiagnosUs (Press 2021), an online medical data annotation platform, regarding whether images of white blood cells contain blast cells. The data from Study 1 are used to create crowdsourced data sets and train the ML models evaluated in this work. In Study 2, we analyze binary classification plus confidence judgments from expert annotators (pathologists and laboratory professional) on the same task and use their responses to evaluate the similarity between expert judgments and the ML models trained on the crowdsourced data sets. Together, these studies allow us to evaluate our methods in terms of AI decision aid accuracy, calibration, and alignment with expert judgments.

2. Study 1

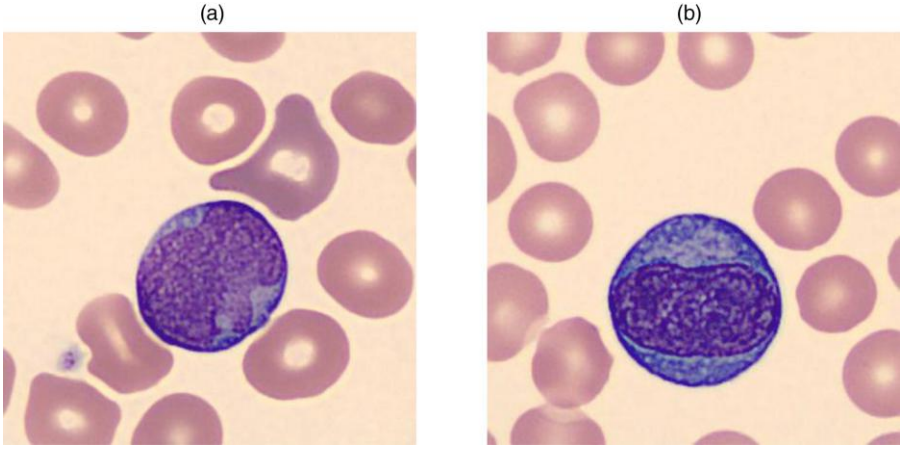
The data examined in Study 1 were also reported by Epping et al. (2026) in Experiment 2. In this study, participants were shown images of white blood cells and asked to judge the likelihood that each image contained a blast cell. This study had two conditions: a binary choice (BC) condition and a probability judgment (or elicited belief (EB)) condition. In the BC condition, participants were shown images of white blood cells and asked whether each image contained a blast cell. In the EB condition, participants were shown images of white blood cells and asked to judge the likelihood that each image contained a blast cell. The previous work evaluated how recalibrating annotators' responses (individual recalibration) impacted the accuracy of both crowdsourced data sets and the ML models trained on the hard label versions of these data sets. The present work builds on the existing work by evaluating (1) how recalibrating crowdsourced labels (crowd recalibration) impacts the accuracy and calibration of crowdsourced data sets and (2) how training ML models on the soft label data sets impacts the accuracy and calibration of the resulting ML models. Below is an overview of the methods for Study 1, and we refer the reader to Epping et al. (2026) for further details.

2.1. Behavioral Methods

The stimuli used in Study 1 were 549 digital images of Wright-stained white blood cells taken from anonymized patient peripheral blood smears at Vanderbilt University Medical Center (VUMC) and were supplied with ground truth labels from three hematopathology faculty from the Department of Pathology, Microbiology, and Immunology at VUMC (Trueblood et al. 2018). Of the 549 white blood cell images, 266 were labeled as blast (cancerous) cells and 283 were labeled as nonblast (noncancerous) cells. Example blast and nonblast cell images are depicted in Figure 1.

On the DiagnosUs app-based crowdsourcing platform, annotation tasks include two types of cases: gold standard (GS) and unlabeled (UL). GS cases have

Figure 1. (Color online) Example White Blood Cell Images



Notes. (a) Example blast cell. (b) Example nonblast cell.

known ground truth and are used to score participants and provide feedback. UL cases have unknown labels, and crowdsourcing is used to generate new annotations. Since the ground truth is unknown, participants do not receive feedback on these UL cases.

In this study, the white blood cell images were randomly divided into a UL set (300 images: 150 blast, 150 nonblast) and a GS set (249 images: 116 blast, 133 nonblast). Although both sets have ground truth labels established by the three hematopathology faculty at VUMC, only GS images are treated as having ground truth during the annotation task. This GS/QA split follows DiagnosUs's standard practice: Participants are scored and receive feedback on GS images, which determine their leaderboard ranking for the task. Participants do not know ahead of time which images are GS images as opposed to UL images. Top performers are eligible for monetary rewards, reinforcing both learning and competition.

A total of 175 participants were recruited through DiagnosUs for Study 1. To prevent participants from competing in both conditions, the app's user base was randomly divided so that half saw only the EB version of the contest and the other half saw only the BC version. A total of 97 users participated in the BC condition, and 78 users participated in the EB condition.

Both conditions consisted of two phases: a practice phase and a competition phase. For practice trials in the BC condition, participants were asked "Do you think this image is a blast cell?" and would respond either "yes" or "no" and were provided feedback after each response. For practice trials in the EB condition, participants were asked "What is the likelihood that this is a blast cell?" and would respond with a probability judgment using a slider and were provided feedback after each response. After completing the practice phase, participants could complete as many competition trials as they would like, but were required to complete at least 200 trials to be eligible for the contest prize. Participants were informed that they were required to complete at least 200 trials to be eligible for the contest prize. Data from participants who failed to completed at least 200 trials and were not eligible for the contest prize were considered incomplete and excluded from the analysis.

The answer prompts in the competition phase were the same as those used in the practice phase. Images in the competition trials were randomly sampled such that one-third of the images were sampled from the GS set, two-thirds of the images were sampled from the UL set, it was equally likely for a blast or nonblast cell to be sampled on each trial, and no images were sampled twice within a 24-hour period. Participants were informed before beginning the contest that blast and nonblast cell would be sampled evenly. Participants received feedback on their responses for images in the GS set but did not receive feedback on their responses for images in the UL set. Because the contests ran over a 48-hour period, it was possible for a single person to see the same image more than once.

To avoid any confounds due to memory effects and ensure each user contributed at most one judgment to each crowdsourced label, we only included a participant's first response in the data analysis if they saw an image more than once.

2.2. Modeling Methods

2.2.1. Recalibration Methods. Both participants' responses and crowdsourced labels were recalibrated using the LLO function (Tversky and Fox 1995, Birnbaum and McIntosh 1996, Turner et al. 2014). As shown in Equation (1), the LLO function recalibrates probabilities such that the transformed probabilities, $f(p)$, are more in-line with objective truth compared with the untransformed probabilities, p:

$$\ln\left(\frac{f(p)}{1-f(p)}\right) = \alpha \ln\left(\frac{p}{1-p}\right) + \beta. \quad (1)$$

Solving for $f(p)$,

$$f(p|\alpha,\beta) = \frac{e^{\beta}p^{\alpha}}{e^{\beta}p^{\alpha} + (1-p)^{\alpha}}. \quad (2)$$

The LLO function has two free parameters: α and β. For individual recalibration, the two free parameters are estimated via maximum likelihood using a participant's responses for images in the GS set. This approach takes advantage of DiagnosUs's existing setup, where GS images have ground truth, so no additional holdout set is needed for estimating recalibration parameters. Letting X_j indicate whether the jth image the participant saw in the GS set was a blast cell ($X_j = 1$) or a nonblast cell ($X_j = 0$) and p_j being the participant's probability judgment that this image was a blast cell, the likelihood function is

$$L(\alpha,\beta|X) = \prod_i f(p_j|\alpha,\beta)^{X_j}[1 - f(p_j|\alpha,\beta)]^{1-X_j}. \quad (3)$$

The slope parameter (α) impacts the curvature of the LLO function and aims to counteract over/underextremity biases, whereas the intercept parameter (β) impacts the height of the LLO function and aims to counteract over/underprediction biases. When $\alpha = 1$ and $\beta = 0$, the LLO function reduces to the identity function. When $\alpha = 0$ and $\beta = 0$, the LLO function maps all input, p, to 50%.

After the two free parameters are fit using a participant's responses for images in the GS set, the LLO function can then be used to transform all of the participant's responses, both for images in the GS set and UL set. Note, individual calibration was carried out independently for each participant because there can be large individual differences in the biases that lead to miscalibration across annotators (Baron et al. 2014), so the α and β parameters of the LLO function vary across participants. We will refer to participants' subjective probability judgments as EB (elicited belief) judgments and participants' recalibrated subjective probability judgments as recalibrated EB (rEB) judgments.

2.2.2. Crowdsourced Label Generation. For EB and rEB data sets, the crowdsourced label for a given image was generated by taking the mean of nine randomly sampled judgments for that image. For BC data sets, the crowdsourced label for a given image was generated by computing the proportion of "blast" judgments from nine randomly sampled judgments for that image (e.g., if four blast judgments were sampled and five nonblast judgments were sampled, the crowdsourced label would be 44.4% blast). Typically, crowdsourced label accuracy increases asymptotically as the number of judgments per label increases (Fiechter and Kornell 2021). The upper limit on accuracy and the rate at which accuracy approaches this limit is dependent on the task, the crowd (in terms of both individual annotator accuracy and diversity across annotators), and sophistication of the aggregation algorithm (Davis-Stober et al. 2014, Lee and Lee 2017, Hasan et al. 2024a). Given that it costs money to collect more judgments per image and gains in accuracy decrease as the number of judgments increases, real-world crowdsourcing tasks aim to strike a balance between maximizing accuracy and minimizing cost. We chose to generate each crowd label using only nine judgments because, anecdotally, real-word medical data annotation tasks rarely collect more than 10 annotations per image (Duhaime et al. 2023).

This process of randomly sampling nine judgments and taking the mean to obtain the crowdsourced label for a given image was repeated 100 times for each image to generate 100 BC, EB, and rEB crowdsourced data sets, with each data set consisting of 549 labels. For crowd recalibration, the LLO function was fit to the 249 crowd labels for images in the GS set and then used to transform the 300 crowd labels for images in the UL set. Similar to how individual recalibration

was carried out independently for each participant, crowd recalibration was carried out independently for each data set. Given that we wanted to evaluate the impact of both individual and crowd recalibration, there were six different types of crowdsourced data sets: BC without crowd recalibration, BC with crowd recalibration, EB without crowd recalibration, EB with crowd recalibration, rEB without crowd recalibration, and rEB with crowd recalibration.

2.2.3. ML Methods. The ML model training/testing and hyperparameter selection was similar to that of Epping et al. (2026) and Holmes et al. (2020). That is, we took a GoogLeNet deep convolutional neural network (Szegedy et al. 2015) pretrained on the ImageNet database and applied transfer learning to fine-tune the model to classify white blood cells as either blast or nonblast cells. To create the hard label data sets from the crowdsourced data sets, we binarized the crowdsourced labels at 50%. That is, if the crowdsourced label was <50%, then the hard label was set to zero (a nonblast label); if the crowdsourced label was >50%, then the hard label was set to one (a blast label); if the crowdsourced label was equal to 50%, it was randomly set to either zero or one. The soft label data sets were the same as the crowdsourced data sets. Like the previous work, only images from the UL set were used for ML model training/testing and hyperparameter selection. The 300 images were divided into training and testing splits using fivefold cross-validation such that each model is trained using approximately 80% of the images and evaluated using approximately 20% of the images, with the constraint that each image appeared in the testing split exactly once across the five folds and the proportion of blast cells in the training and testing splits for each fold were roughly equal. Hyperparameter selection was carried out using a grid search over potential hyperparameter values for both hard and soft label data sets for each of the six different types of labeled data sets. Therefore, the hyperparameter selection process was carried out 12 times. Here, the grid search was over the number of training epochs (5, 10, 15, 20, and 25), learning rate ($1e^{-4}$, $5e^{-4}$, and $1e^{-3}$), L2-regularization strength ($1e^{-4}$, $5e^{-4}$, $1e^{-3}$, $5e^{-3}$, and $1e^{-2}$), and minibatch size (16, 24, and 32). The hyperparameters of the models that yielded the lowest binary cross-entropy (BCE) loss, on average, between the models' output and the crowd labels on the testing splits were deemed optimal. The results for the entire grid search over hyperparameter values can be found in the supplement. After the optimal hyperparameters were identified, the model training/testing procedure was carried out using the optimal hyperparameters, where the models were still trained on the crowd labels, but now evaluated using the ground truth labels. The fivefold cross-validation process was repeated 20 times, yielding 100 models such that each model was trained/evaluated on a different data set.

2.2.4. Summary. From this raw data, we produced 12 different processed data sets for model training. These data sets differ on three factors.

1. Labeling approach (BC, EB, rEB). Data were collected as either binary labels (BC) or label probabilities (EB). The EB labels can then be left as reported or recalibrated (rEB) at the individual level.

2. Crowd recalibration (present or absent). Individual level labels are averaged into crowd labels prior to training. These crowd labels can be either recalibrated (present) or not (absent).

3. Training label type (hard, soft). The resulting crowd level labels can either be binarized (hard) or left in their probabilistic form in the interval [0, 1] (soft).

Crossing these three factors yields 12 different sets of labels for the image bank. The ML model was trained separately on each of these twelve different processed forms of the data. For each of these 12 data variants, model training is performed using fivefold cross-validation (varying test/train split) is performed 20 times (varying the nine individual labels drawn from the full data set for each image), resulting in 100 trained models for each data variant.

2.3. Behavioral Results

Prior to ML model training, we first measure each crowdsourced data set's accuracy and calibration, which were evaluated only using images from the UL set, because the GS set was used to fit the recalibration model parameters. When evaluating accuracy, crowd labels >50% are treated as blast labels and crowd labels <50% are treated as nonblast labels when evaluating accuracy. If the crowd label was equal to 50%, then the crowd labels was regarded as being partially correct (accuracy = 0.5) regardless of the true label.

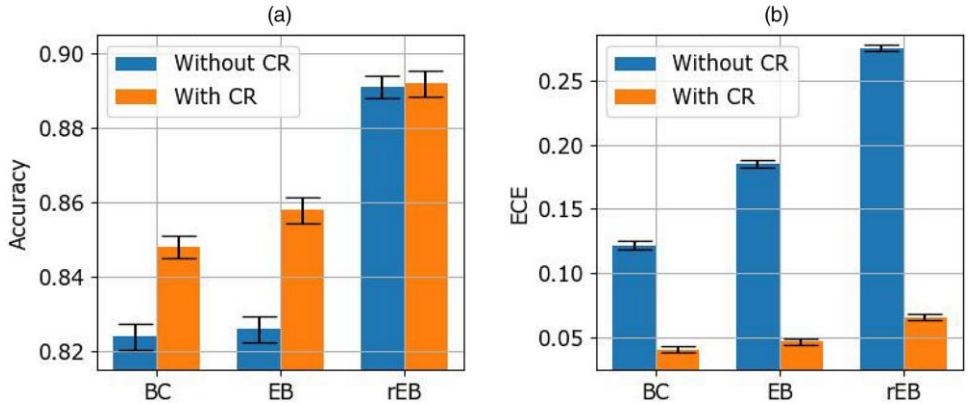
Each data set's calibration was measured using expected calibration error (ECE) (Guo et al. 2017). ECE is computed by first dividing the crowd labels for a single data set into N equally spaced bins (here we used $N = 10$). The calibration error for a given bin is equal to the distance between the average label in that bin and the proportion of blast cell images in that bin. ECE is then defined as the weighted sum of the calibration errors across bins, where the weight for a given bin is equal to the number of images in that bin divided by the total number of images in the data set (300). In the supplement, we included a similar plot depicting the mean squared error (MSE) of the various data sets because MSE accounts for both accuracy and calibration in a single measure. The rEB data sets with crowd recalibration had the lowest MSE. Figure 2 contains the accuracy and ECE results for the six different types of labeled data sets. Figures illustrating the 100 pairs of α and β parameters that result from performing crowd recalibration on the BC, EB, and rEB data sets can be found in the supplement.

Note, ECE is a measure of error, so a lower ECE value indicates a better calibrated data set. Based on Figure 2, the most accurate data sets are the rEB data sets, both with and without crowd recalibration, and the best calibrated data sets are the BC and EB data sets with crowd recalibration. Although individual recalibration leads to improvements in accuracy, it worsens ECE. Crowd recalibration improves the accuracy of the BC and EB data sets, in addition to improving the ECE for BC, EB, and rEB data sets. Figure 3 illustrates the impact of crowd recalibration on the accuracy and calibration of the 100 BC, EB, and rEB data sets.

In Figure 3, (a) and (b), nearly all of the points lie above the $y = x$ line, indicating that crowd recalibration consistently improved the accuracy of BC and EB data sets. In Figure 3(c), there are roughly an even number of points above and below the $y = x$ line, indicating that crowd recalibration did not have a notable, consistent impact on the accuracy of rEB data sets. In Figure 3, (d), (e), and (f), all of the points lie well below the $y = x$ line, especially for the rEB data sets, indicating that crowd recalibration always improved the calibration of BC, EB, and rEB data sets.

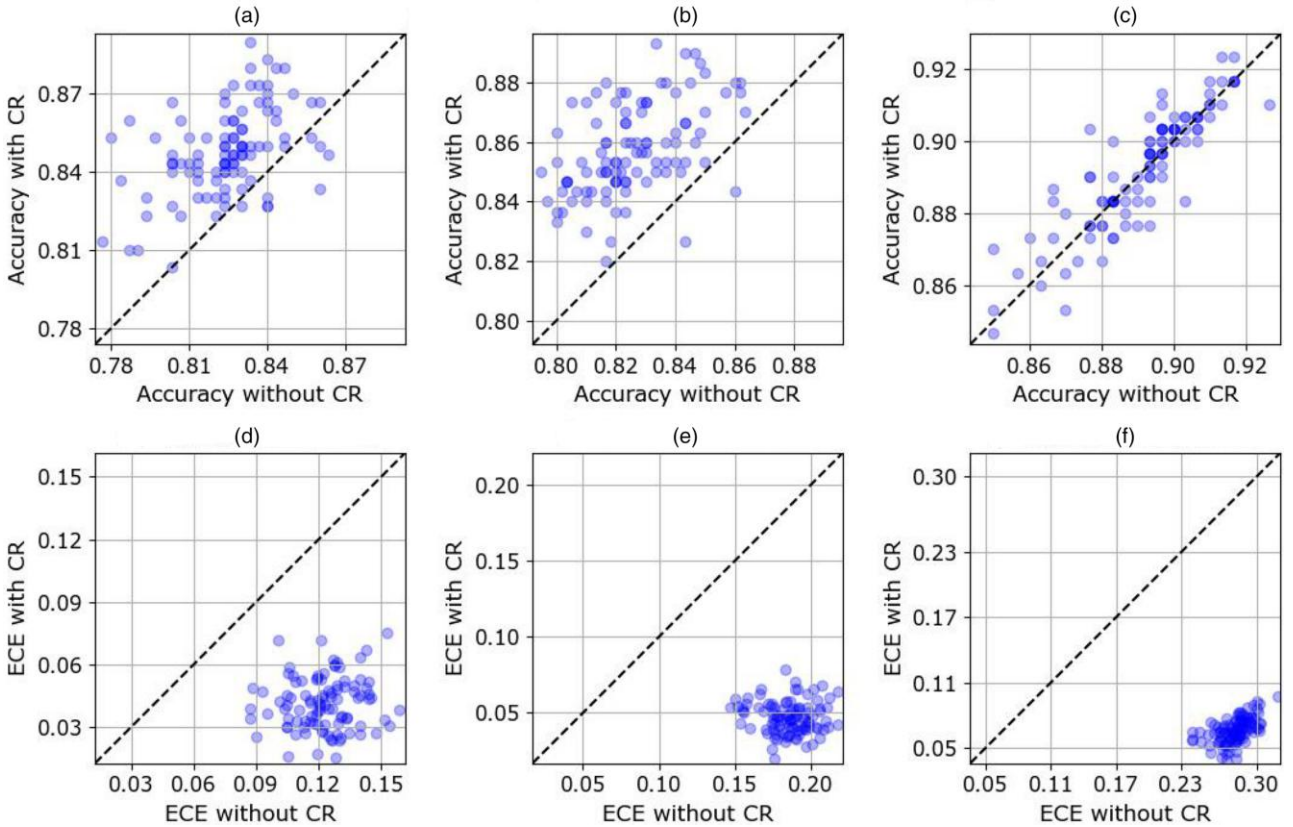
To better understand why crowd recalibration impacted the accuracy and calibration of BC and EB data sets differently compared with how crowd recalibration impacted that of rEB data sets, we plotted the calibration curves for one of the BC, EB, and rEB data sets before and after crowd recalibration in Figure 4. Calibration curves are generated by binning labels into seven equally spaced intervals. Although the number of bins is arbitrary, we chose seven because it best illustrates the effects of crowd recalibration. In the figure, there is one point for each bin that contains at least two labels. The x value for each point represents the mean label across all labels in the bin

Figure 2. (Color online) Labeled Data Set Results



Notes. (a) Accuracy for the six types of labeled data sets. Error bars represent the 95% confidence intervals for the mean accuracy across the 100 data sets for each type of data set. (b) ECE (computed using 10 bins) for the six types of labeled data sets. Error bars represent the 95% confidence intervals for the mean ECE across the 100 data sets for each type of data set. CR, crowd recalibration.

Figure 3. (Color online) Impact of Crowd Recalibration

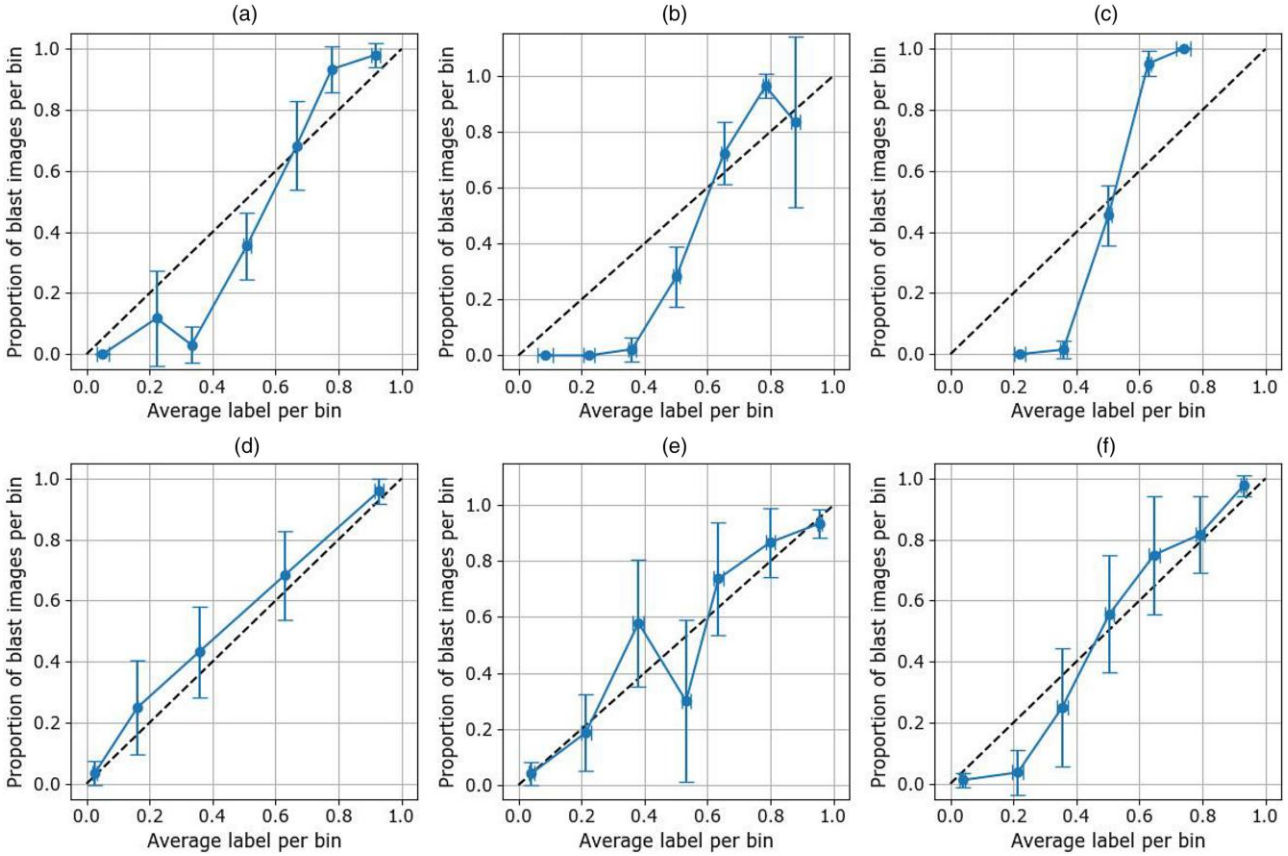


Notes. Accuracy with and without crowd recalibration for (a) BC data sets, (b) EB data sets, and (c) rEB data sets. ECE with and without crowd recalibration for (d) BC data sets, (e) EB data sets, and (f) rEB data sets. CR, crowd recalibration.

corresponding to that point, and the x error bars represent the 95% CI for the mean label in that bin. The y value for each point represents the proportion of blast images in that bin, and the y error bars represent the 95% CI for the proportion of blast images in that bin. For example, the left-most point of Figure 4(a) has coordinates $(0.051, 0)$, indicating that the mean label across all labels that fall within $[0, .143)$ is 0.051 and the true label is blast for 0% of the images to which these labels correspond. When the x value is equal to the y value for a given point, the mean label for that bin is equal to the proportion of blast cells in that bin. Therefore, the closer the calibration curve is to the $y = x$ line, the better the calibration of the data set. The data sets shown in the figure were selected because they were the most prototypical data sets out of the 100 for each data set type. By prototypical, we mean that the accuracy and ECE values before and after crowd recalibration were the closest to the mean accuracy and ECE values before and after crowd recalibration across all 100 data sets.

We see that the calibration curves for all three data sets without crowd calibration indicate an underextremity bias, although to a greater extent for the rEB data set. That is, the crowd labels below 50% overestimate the likelihood of blast and the crowd labels above 50% underestimate the likelihood of blast. This trend is corroborated by the best fit α parameters from the LLO functions fit to the 100 BC, EB, and rEB data sets. Note that $\alpha > 1$ indicates an underextremity bias. The mean best fit α parameter is 2.12 for the BC data sets, 3.39 for the EB data sets, and 4.88 for the rEB data sets, confirming that all three data sets without crowd calibration exhibited an underextremity bias. The rEB data sets exhibited this bias to the greatest extent. Additionally, the calibration curves for the BC and EB data sets

Figure 4. (Color online) Impact of Crowd Recalibration on Calibration Curves for Example BC, EB, and rEB Data Sets



Notes. (a) Calibration curve for an example BC data set without crowd recalibration. (b) Calibration curve for an example EB data set without crowd recalibration. (c) Calibration curve for an example rEB data set without crowd recalibration. (d) Calibration curve for an example BC data set with crowd recalibration. (e) Calibration curve for an example EB data set with crowd recalibration. (f) Calibration curve for an example rEB data set with crowd recalibration. CR, crowd recalibration.

exhibit an overprediction bias, where the majority of the points on the calibration curves fall under the $y = x$ line. This trend is also corroborated by the best fit β parameters from the LLO functions fit to the 100 BC, EB, and rEB data sets. Note that $\beta < 0$ indicates an overprediction bias. The mean best fit β parameter is -0.915 for the BC data sets, -1.057 for the EB data sets, and -0.157 for the rEB data sets, confirming that the BC and EB data sets exhibit an overprediction bias. The results also suggest that the rEB data sets exhibit this same bias, but to a much lesser degree.

Crowd recalibration brings the calibration curve closer to the $y = x$ line for all three data sets, suggesting that crowd recalibration corrects for nearly all these biases. Correcting for an underextremity bias does not impact the accuracy of the labeled data set because this correction involves making the distribution of labels more extreme, without shifting many probabilities from <50% to >50% or vice versa. Therefore, crowd recalibration greatly improves the ECE of the rEB data sets without having a significant impact on accuracy by correcting for an underextremity bias. Unlike correcting for underextremity, correcting for overestimation bias can lead to significant improvements in data set accuracy by shifting a large number of labels from <50% to >50%. Figure 2 supports this notion, because BC and EB data sets with crowd recalibration are more accurate compared with BC and EB data sets without crowd recalibration. Therefore, crowd recalibration leads to improvements in BC and EB data set calibration and accuracy by correcting both underextremity and overprediction biases and

leads to improvements in rEB data set calibration without impacting accuracy by correcting an underextremity bias.

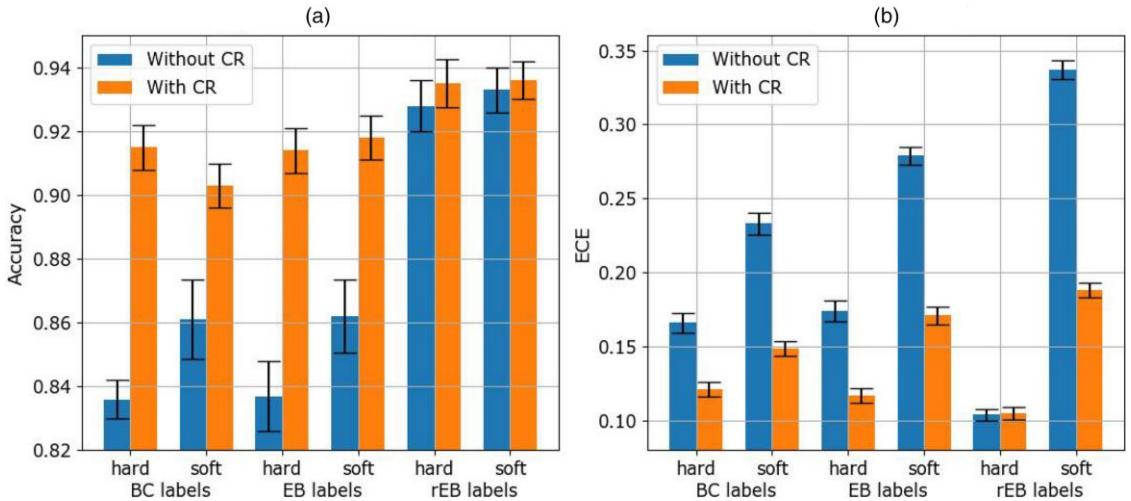
2.4. ML Results

The accuracy and calibration of the ML models were measured the same way we measured the labeled data sets, except that the ML models were only evaluated using their output on the testing set. Figure 5 contains the accuracy and ECE results for all 12 ML model variants.

Based on Figure 5, the most accurate models were trained on the rEB data sets, regardless of whether the labels were transformed via crowd recalibration or the models were trained on hard or soft labels. The best calibrated models are those trained on the rEB hard label data sets, both with and without crowd recalibration. Crowd recalibration had little impact for models trained on hard labels because it only alters a small subset of labels (those are moved from <50% to >50% and vice versa) due to the hard labels being binarized at 50%. For models trained on soft labels, crowd recalibration led to models that were better calibrated and more accurate (at least for models trained on BC and EB data sets) compared with those trained on data sets without crowd recalibration. Training on soft labels appeared to have either a positive or negligible impact on accuracy, but always led to a worsening of ECE.

Training on soft labels led to a worsening of ECE because of the loss function used during model training, in conjunction with the fact that the soft labels either exhibited an underextremity bias (without crowd recalibration) or a negligible over/underextremity biases (with crowd recalibration). With soft labels, the loss function penalizes models more for overconfident judgments compared with underconfident ones. For example, if the soft label for a given image is 80% blast, models are penalized more strongly if they output 90% blast compared with 70% for that image. As a result, the models are trained to err on the more conservative side and output judgments that are more conservative relative to the soft labels on which they are trained. Therefore, because soft labels without crowd recalibration already exhibit an underextremity bias, this bias becomes more pronounced in the models trained on those labels. Similarly, because soft labels with crowd recalibration exhibit negligible over/underextremity biases, an underextremity bias emerges in the models trained on those labels. In either case, the models will have a greater underextremity bias compared with the data they are trained on, leading to a worsening of ECE. Models trained on hard labels are not affected in the same way because they produce very extreme judgments, which results in ECE $\approx 1 -$ accuracy. This relationship between ECE

Figure 5. (Color online) Labeled Data Set Results



Notes. (a) Test accuracy for the ML models trained on the 12 data variants. Error bars represent the 95% confidence intervals for the mean accuracy across the 100 models for each data variant. (b) ECE (computed using 10 bins) for the ML models trained on the 12 data variants. Errorbars represent the 95% confidence intervals for the mean ECE across the 100 models for each data variant. CR, crowd recalibration.

and accuracy for models trained on hard labels can be seen in Figure 5.

2.5. Conclusion

In Study 1, we found that individual recalibration led to improvements in accuracy for both crowdsourced data sets and the ML models trained on these data sets. Crowd recalibration led to improvements in accuracy and calibration for both BC and EB crowdsourced data sets and the models trained on these data sets by correcting for simultaneous underextremity and overprediction biases. Also, crowd recalibration led to improvements in calibration for rEB data sets and models trained on the rEB soft label data sets by correcting for an underextremity bias. Training models on soft labels led to minor gains in accuracy, for the most part, and worsening of calibration. Our results from Study 1 offer strong evidence in support of both individual and crowd recalibration, but not for training models on soft labels. Recall, our approach involves improving model accuracy, calibration, and alignment with expert judgments. Although soft labels generally led to worse ML calibration than hard labels, we hypothesized that they might produce models that are better aligned with expert judgments. Accordingly, Study 2 evaluates this hypothesis by comparing ML model outputs to expert judgments across different training data variants.

3. Study 2

The data examined in Study 2 were reported by Hasan et al. (2024b) in Experiment 2. In this study, expert participants made classification decisions about white blood cells in two parts. In the first part, participants were asked whether images contained a *blast* cell and then asked to estimate their confidence that their judgment was correct. In the second part, they were asked whether the same images contained a *nonblast* cell and again asked to provide a confidence judgment. Hasan et al. (2024b) evaluated the accuracy of Wisdom of the Crowd algorithms when only including participant's most confident judgment per image in the crowd decision. In the present work, we evaluate the similarity between the output of the ML models trained in Section 2.4 and the expert responses from the first part of this study. Below is an overview of the methods for Study 2, and we refer the reader to Hasan et al. (2024b) for further details.

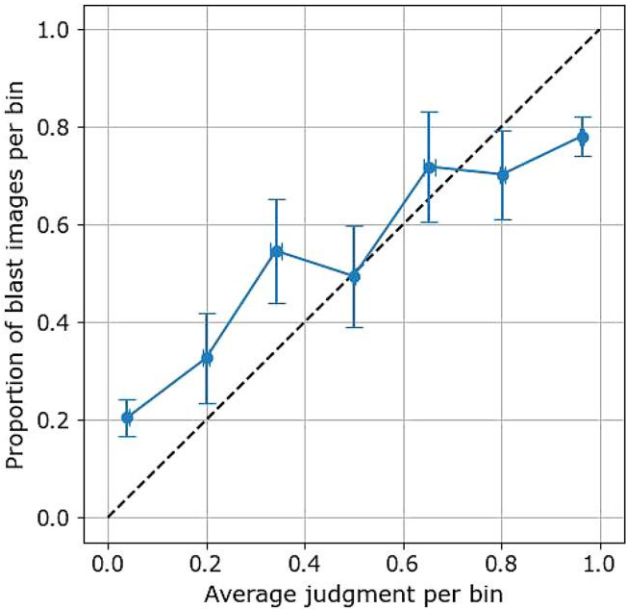
3.1. Behavioral Methods

For this study, 22 pathologists and laboratory professionals who had experience classifying white blood cells were recruited at the American Society for Clinical Pathology conference. We will refer to the participants in Study 2 as experts given their prior experience. Study 2 consisted of two phases: a practice phase and a testing phase. During the practice phase, participants were shown a pair of images on each trial (one blast and one nonblast) and a label (either "blast" or "nonblast"). They were asked to select which image matched the label, and were given feedback after their response. The testing phase was divided into two parts with 60 trials in each part. During the first part, participants were shown an image of a white blood cell and asked "Is this a blast cell?" and then they provided a confidence judgment in their response on a scale from 0 to 50. During the second part, participants were shown an image of a white blood cell and asked "Is this a nonblast cell?," and then they provided a confidence judgment in their response on a scale from 0 to 50. The same 60 images (30 blast images and 30 nonblast images) were used in both parts of the testing phase, the only difference being that the images in the second part were rotated 180°. All 22 experts saw the same set of images in the testing phase and these images were a subset of the 549 images used in Study 1. Note, these images were specifically chosen because the three hematopathology faculty at VUMC who provided ground truth labels (Trueblood et al. 2018) deemed these 60 images difficult to correctly classify, and Hasan et al. (2024b) wanted to use difficult images to challenge experts so that they would express uncertainty in their decisions. We only used responses for the first part of the testing phase because the prompt "Is this a blast cell?" more closely aligns with the prompts used in Study 1.

3.2. Expert-Model Similarity Results

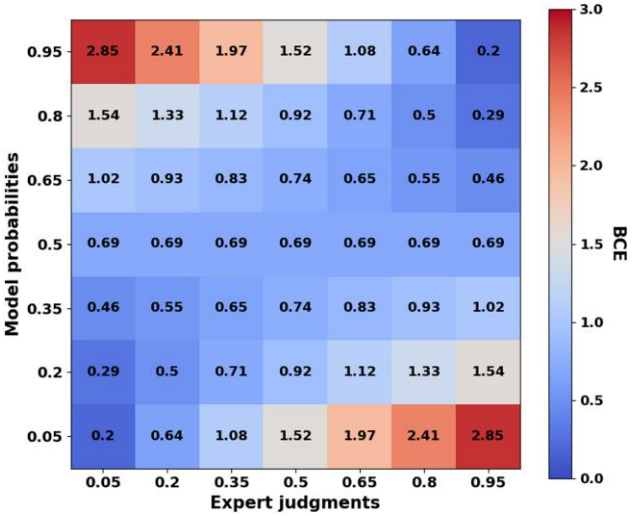
To evaluate the similarity between the ML models and the expert judgments, we first needed to transform the experts' classification plus confidence judgments into a probability judgment. If the expert responded "Yes" to the prompt "Is this a blast cell?," we defined their probability judgment as being equal to 50% plus their confidence judgment. If the expert responded "No" to the same prompt, we defined their probability

Figure 6. (Color online) Expert Judgments Calibration Curve



Note. Calibration curve for judgments pooled across experts.

Figure 7. (Color online) BCE for a Given Expert Judgment and Model Probability



Note. Heatmap showing BCE values for combinations of expert judgments and model probabilities, with judgments/probabilities ranging from 0.05 to 0.95.

judgment as being equal to 50% minus their confidence judgment. Figure 6 illustrates the calibration curve for the expert judgments. Based on the figure, we see that the expert judgments exhibit an overextremity bias and a slight overprediction bias.

To compare the similarity of a single ML model to a single expert, we paired each expert with each ML model. Because 100 ML models were trained for each data variant and there were 22 experts, there were a total of 2,200 expert-model pairs for each data variant. With each expert-model pair, the similarity in their judgment was defined as the mean BCE between the expert's judgments and the output from the ML model for a given data set (Xu and Xia 2012). Figure 7 presents a heatmap designed to help interpret BCE values by showing how they vary across different combinations of expert judgments and model probabilities. As shown in the figure, when both the expert judgment and model probability agree and are highly confident (e.g., both are 0.05 or both are 0.95), BCE values are around 0.2. However, in cases of strong disagreement, such as when the expert judgment is 0.05 and the model probability is 0.95, the BCE is around 2.85. Thus, large BCE values reflect larger disagreement between experts and models.

BCE can only be computed for images that received judgments from both the expert and ML model, so the similarity for each expert-model pair was only computed over the images that were among the 60 seen by the given expert and were included in the testing split for the given ML model. This results in similarity being computed using an average of 5.98 judgments for each expert-model pair. Although this results in a low number of judgments used to evaluate the similarity for each expert-model pair, we chose to include these 60 images in the ML model training/testing for two reasons. First, because there are 300 images used for ML model training/testing in Study 1 and the images are split such that 80% are used for training and 20% are used for testing, we are training those ML models using only 240 images, which is already a very small training set. Had we held out these 60 images from ML model training, the ML training set would have only consisted of 192 images, making training even more difficult. Second, as stated in Section 3.1, the 60 images used in Study 2 were selected because they were judged to be difficult to correctly classify. Thus, the set of 60 images in this study likely contained different characteristics compared with the remaining 489 images. ML practitioners aim to curate training data sets that accurately reflect the characteristics of

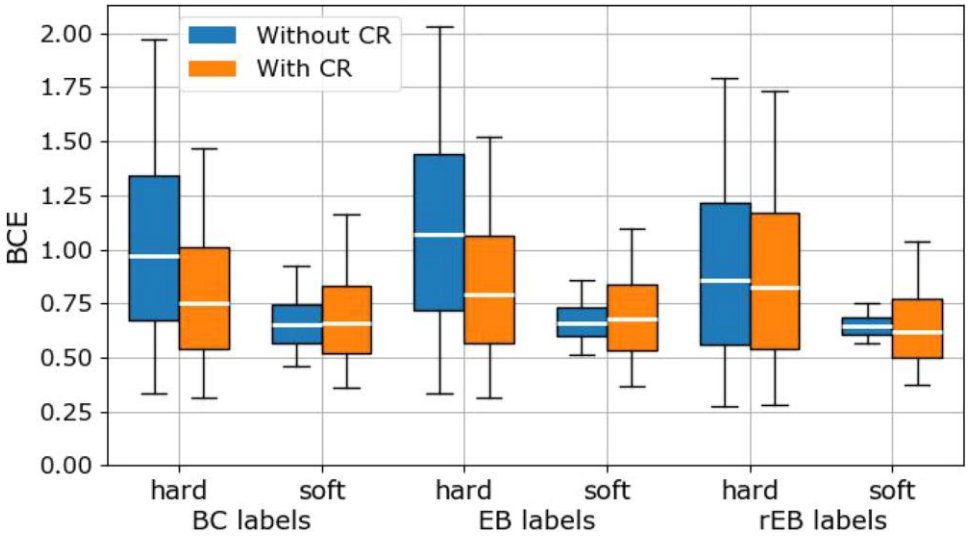
the out-of-sample data set because ML classifiers can only learn to classify based on the statistical regularities in their training data set. ML models will likely perform poorly on a testing data set if the data in the testing data set has different characteristics compared with the data in the training data set. Therefore, we chose to include these 60 images in the ML model training/testing to align our training/testing procedures with those typically used in real-world ML applications.

Figure 8 displays boxplots that describe the BCE distribution across the 2,200 expert-model pairs for each data variant. Like ECE, BCE is also a measure of error, so lower BCE values in Figure 8 correspond to higher expert-model similarity. We found that models trained on soft labels output judgments that were more similar to experts compared with models trained on hard labels, regardless of whether they were trained on BC, EB, or rEB data sets with or without crowd recalibration. Additionally, crowd recalibration appears to improve the similarity between experts and models for models trained on the BC and EB hard label data sets.

We also see that the variance of the BCE distributions for all models trained on hard labels is greater than that for all models trained on soft labels. This makes sense given that the models trained on hard labels are more likely to produce extreme judgments compared with models trained on soft labels. Models trained on hard labels are more likely to produce extreme judgments because, during training, loss is minimized by aligning the ML models' output with its training labels. In order to align its output with hard labels, the ML model learns to output judgments close to 0% and 100%. Therefore, when a model trained on hard labels agrees with an expert's classification (either both the model and expert have probability judgments greater than 50% or both have judgments less than 50%), their probability judgments are relatively similar. However, when the two disagree, their probability judgments are very dissimilar. Because the models trained on soft labels are more conservative in their probability judgments, the expert and model probability judgments can still be relatively similar even when the two disagree on the correct classification. Therefore, models trained on hard labels tend to produce more extreme judgments, which increases the chance of strong disagreement with experts and leads to greater variance and longer tails in the BCE distributions, as indicated in Figure 8.

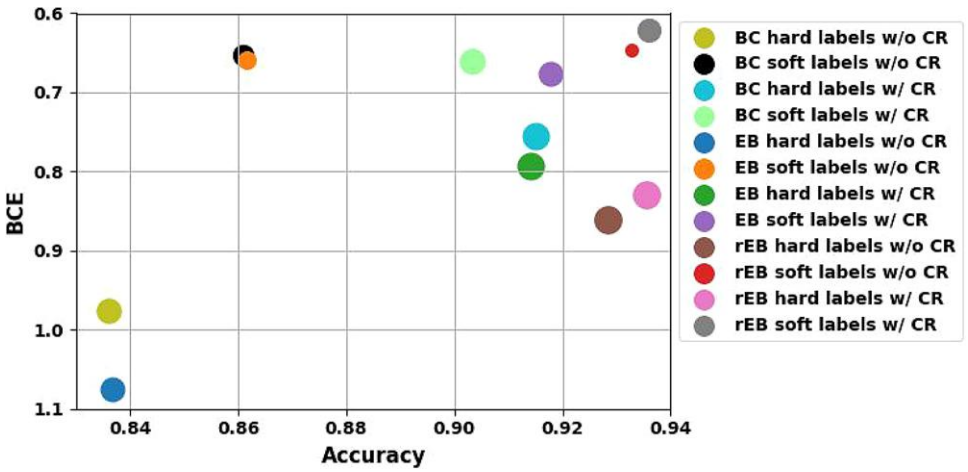
Crowd recalibration appears to shift the BCE distributions toward lower BCE values for models trained

Figure 8. (Color online) Expert-Model Similarity



Notes. Expert-model similarity for ML models trained on 12 data variants. For each variant, the white line represents the median BCE, the lower limit on the box represent the first quartile (25th percentile BCE), the whisker extending from the lower limit on the box represents the 5th percentile BCE, the upper limit on the box represents the third quartile (75th percentile BCE), and the whisker extending from the upper limit on the box represents the 95th percentile BCE. CR, crowd recalibration.

Figure 9. (Color online) All ML Results



Notes. Each point corresponds to an ML model trained on one of the 12 data variants. The x value is the mean testing accuracy across the 100 ML models trained on a particular data variant. The y value is the median BCE across the 2,200 expert-model pairs for the corresponding data variant. The size of the point is inversely proportional to the mean ECE at test for the 100 ML models trained on the data variant. CR, crowd recalibration.

on BC and EB hard label data sets. This is likely due to crowd recalibration improving the accuracy of models by correcting for the overprediction bias in the crowd-sourced data sets. By improving the accuracy of the models, crowd recalibration lowers the likelihood of a model outputting a highly confident, incorrect judgment and, as a result, there are fewer expert-model pairs that strongly disagree with one another. Note, we also measured the similarity between the expert's binary classifications and the model outputs using the same procedure outlined here and included a similar boxplot figure in the Online Appendix. The results did not change, aside from the box edges and whiskers expanding.

Figure 9 summarizes the performance of the ML models for each data variant across the tricriteria (accuracy, calibration, and alignment with expert judgments) and highlights tradeoffs among these metrics. The ML models trained on the rEB soft label data sets with crowd recalibration are the most accurate and most similar to experts, closely followed by the ML models trained on the rEB soft label data sets without crowd recalibration. Nonetheless, these two models do not have the best ECE (as indicated by the size of their corresponding points). The best ML model variants in terms of ECE are those trained on the rEB hard label data sets. These models are also roughly as accurate as those trained on the rEB soft label data sets, but this improvement in ECE comes at the cost of a decrease in similarity with experts. Note, the ML models that produced the lowest accuracy and the lowest agreement with experts are the models trained on BC and EB hard labels without crowd recalibration, which were the only two models trained without any of our proposed methods (individual recalibration of subjective probability judgments, crowd recalibration, and soft label training). Therefore, each of our proposed methods positively contributed toward our intended goal of improving model accuracy, calibration, and agreement with experts positively.

3.3. Conclusion

In Study 2, we found that models trained on soft labels produced judgments that were more similar to experts compared with models trained on hard labels. Therefore, Study 2 offers strong evidence, lacking in Study 1, for the efficacy of training models on soft labels. We note that there appears to be a tradeoff between expert-model alignment and model calibration. Although ML calibration is often emphasized, alignment with expert judgments may play a critical role in AIADM because it could influence whether users are willing to engage with and trust the AI's advice. Even a perfectly calibrated model may have limited impact if users disregard its input. This raises the possibility that, in some AIADM contexts, alignment may be more important

than calibration. From this perspective, the ML models trained on the rEB soft label data sets may be particularly well suited for AIADM, given their strong alignment with expert judgments.

4. Discussion

This work examined how characteristics of human-annotated data affect ML model accuracy, calibration, and alignment with expert judgments, and explored how correcting systematic miscalibration in these annotations can enhance ML model performance on these criteria. In Study 1, we analyzed binary classifications and probability judgments from a white blood cell classification task in a field experiment conducted on the data annotation app, DiagnosUs. We aggregated responses across annotators using wisdom of the crowd methods to create different crowdsourced data sets for training ML models. We trained ML models on 12 different data variants, defined by three factors: labeling approach (binary choices, elicited beliefs, and recalibrated elicited beliefs), presence or absence of crowd-level recalibration, and label type (hard or soft). A central focus of Study 1 was assessing how recalibrating both individual and crowd-level judgments influenced model accuracy and calibration. In Study 2, we compared expert confidence judgments on the white blood cell task to the outputs of ML models trained on each of the 12 data sets. ML models trained on soft labels from both individual and crowd-recalibrated data were the most accurate and showed the strongest alignment with expert judgments, although these models were not the best calibrated. Our findings highlight that there can be a tradeoff between model calibration and alignment with human judgments. Nevertheless, because model accuracy and alignment with expert judgments are likely to influence users' trust in AI decision aids, models trained on rEB soft label data sets with crowd recalibration may be especially well positioned to support effective AIADM systems.

In Study 1, we used the LLO function to recalibrate probability judgments by fitting two parameters that correct common sources of miscalibration. When judgments are poorly calibrated, the distance from 50% is a weak indicator of classification accuracy. Because confidence-weighted wisdom of the crowd methods rely on the reliability of these probability estimates, improving individual calibration by correcting systematic biases can significantly enhance the accuracy of crowdsourced data sets. In support of this, we found that individual recalibration notably improved crowdsourced data set accuracy, because rEB data sets were more accurate compared with EB data sets in Study 1. In addition to recalibrating individual subjective probability judgments, recalibrating the crowdsourced labels can improve crowdsourced data set accuracy by mitigating over/underprediction biases. In evidence, we found that crowd recalibration improved the accuracy of the BC and EB data sets in Study 1 by counteracting an overprediction bias. Overall, applying the LLO function, whether to individual probability judgments or to crowdsourced labels, led to more accurate data sets, with individual recalibration producing the highest gains in accuracy. In addition to improving accuracy, crowd recalibration also improved the calibration of BC, EB, and rEB data sets in Study 1.

The improvements in crowdsourced data set accuracy carried over to the ML models trained on these data sets. In Study 1, we found that ML models trained on the BC and EB data sets with crowd recalibration and those trained on the rEB data sets were significantly more accurate compared with the models trained on BC and EB data sets without crowd recalibration. Furthermore, ML models trained on BC, EB, and rEB data sets with crowd recalibration were better calibrated than models trained on the same data sets without crowd recalibration. Therefore, we conclude that our proposed solution of recalibrating individual judgments and crowdsourced labels via the LLO function accomplished our stated goal of improving the accuracy and calibration of ML models trained on crowdsourced data sets.

To improve ML model alignment with experts, we proposed to train ML models on soft labels rather than hard labels. Hard labels provide only categorical outcomes and ignore the uncertainty associated with each data instance, whereas soft labels capture this uncertainty. In crowdsourced data sets, such uncertainty naturally arises from aggregating judgments across multiple annotators. Additionally, recalibrating the crowdsourced labels aimed to ensure that the uncertainty reflected in the soft labels was well calibrated. When comparing the models with experts in Study 2, we found that models trained on soft labels exhibited greater alignment with expert judgments compared with those trained on hard labels. Importantly, the models trained on the rEB soft

label data set with crowd recalibration were the most accurate and exhibited the highest similarity to expert judgments, whereas the models trained on the BC and EB hard label data sets without crowd recalibration were the least accurate and exhibited the lowest similarity to expert judgments. Thus, the three proposed techniques (individual recalibration, crowd recalibration, and the use of soft labels) each helped create ML models that achieved our goal of improving model accuracy, calibration, and alignment with expert judgments.

We were surprised to find that crowd recalibration did not consistently improve alignment between expert judgments and models trained on soft labels. Although crowd recalibration reliably improved model calibration, it had less impact on expert-model alignment, likely because the expert judgments themselves were not well calibrated. This is consistent with our finding that the best-calibrated models were not always the most aligned with expert judgments. Alignment between models and experts depends not just on model calibration, but also on the calibration of the expert judgments themselves. Our results point to the need for future research to better understand the properties, such as model calibration and expert-model alignment, that contribute to effective AIADM systems.

4.1. Constraints on Generality

This work has several important limitations. First, we evaluated our methods using only a white blood cell classification task. Although this provides a valuable test case, it is unclear whether the findings will generalize to other domains. Second, we only applied our techniques to a binary classification problem. Extending these methods to handle multiclass classification tasks will be crucial for broader applicability, as many real-world decision-making settings involve multiple possible outcomes. Third, our data set is small compared with the size of data sets normally employed in ML studies and, in particular, our study of model-expert alignment relied on expert responses to a relatively small sample of 60 images. This limited data set may constrain the robustness of our conclusions and highlights the need for future work examining model-expert alignment with larger data sets. Fourth, we limited ourselves to a simple linear averaging algorithm when aggregating individual annotations into crowdsourced labels. The recalibration approach outlined in this work can be combined with more sophisticated algorithms that differentially weight annotators' judgments based on each annotator's past performance when aggregating their judgments into consensus labels (Wang et al. 2011, Budescu and Chen 2015, Collins et al. 2023, Hasan et al. 2024a). Evaluating whether our recalibration approach yields gains in ML model accuracy and expert alignment when combined with these more complex algorithms remains an area for future work.

4.2. Future Directions

The broader goal of our work is to support human-AI complementarity, where an AIADM system outperforms either the human or the AI alone. To fully evaluate the impact of our proposed techniques on human-AI complementarity, the next step is to deploy these models as decision aids in real-world settings. We view the current work as a necessary foundation for such future studies. Building on prior research showing that people quickly lose trust in models after observing them err (Dietvorst et al. 2015) and that they are more likely to integrate AI advice when it aligns with their own judgments (Grgić-Hlača et al. 2022), we focused on developing models that maximize both accuracy and alignment with experts. Deploying and testing these models in AIADM contexts remains an important direction for future work.

5. Conclusion

AIADM aims to combine the strengths of humans and AI to achieve more accurate judgments than either alone. A major challenge in building effective AIADM systems is addressing human biases, both those that affect the quality of ML training data and those that shape how people trust and engage with AI-generated advice. To tackle these issues, we applied insights from decision science to reduce biases in crowdsourced data sets and trained ML models on data that captured human uncertainty about each label. Our results show that ML models trained on recalibrated soft labels achieve higher accuracy and stronger alignment with expert judgments, although this comes at the cost of reduced ML calibration. In sum, this work underscores the value of harnessing human uncertainty in ML training data and emphasizes the importance of accounting for model accuracy, calibration, and alignment with human judgment when designing AI systems.

References

Agarwal N, Moehring A, Rajpurkar P, Salz T (2023) Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, Cambridge, MA.

Baranski JV, Petrusic WM (1998) Probing the locus of confidence judgments: Experiments on the time to determine confidence. *J. Experiment. Psych. Human Perception Performance* 24(3):929–945.

Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH (2014) Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis (Oxford)* 11(2):133–145.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? Elish MC, Issac W, Zemel RS, eds. *Proc. ACM Conf. Fairness Accountability Transparency* (ACM, New York), 610–623.

Birnbaum MH, McIntosh WR (1996) Violations of branch independence in choices between gambles. *Organ. Behav. Human Decision Processing* 67(1):91–110.

Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R, eds. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proc. Adv. Neural Inform. Processing Systems*, vol. 29 (Curran Associates, Inc, Red Hook, NY), 4356–4464.

Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Sci.* 61(2):267–280.

Cacciamani GE, Sanford DI, Chu TN, Kaneko M, Abreu ALDC, Duddalwar V, Gill IS (2023) Is artificial intelligence replacing our radiology stars? Not yet! *Eur. Urology Open Sci.* 48:14–16.

Caplin A, Deming DJ, Li S, Martin DJ, Marx P, Weidmann B, Ye KJ (2024) The ABC's of who benefits from working with AI: Ability, beliefs, and calibration. Technical report, National Bureau of Economic Research, Cambridge, MA.

Collins KM, Bhatt U, Weller A (2022) Eliciting and learning with soft labels from every annotator. Hsu J, Yin M, eds. *Proc. AAAI Conf. Human Comput. Crowdsourcing*, vol. 10 (AAAI Press, Palo Alto, CA), 40–52.

Collins RN, Mandel DR, Budescu DV (2023) Performance-weighted aggregation: Ferreting out wisdom within the crowd. *Judgment in Predictive Analytics* (Springer), 185–214.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1(2):79.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. Huttenlocher D, Medioni G, Rehg J, Essa I, Kang SB, Pollefeys M, eds. *Proc. IEEE Conf. Comput. Vision Pattern Recognition* (IEEE, Los Alamitos, CA), 248–255.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. General* 144(1):114.

Duhaime EP, Jin M, Moulton T, Weber J, Kurtansky NR, Halpern A, Rotemberg V (2023) Nonexpert crowds outperform expert individuals in diagnostic accuracy on a skin lesion diagnosis task. Lepore N, Acosta O, eds. *Proc. IEEE 20th Internat. Sympos. Biomedical Imaging* (IEEE, Piscataway, NJ), 1–5.

Epping GP, Caplin A, Duhaime E, Holmes WR, Martin D, Trueblood JS (2026) Improving crowdsourcing for AI through cognitive-inspired data engineering. Preprint, submitted December 30, https://osf.io/preprints/psyarxiv/euk26_v1.

Fiechter JL, Kornell N (2021) How the wisdom of crowds, and of the crowd within, are affected by expertise. *Cognition Res. Principle Implications* 6(1):7.

Galton F (1907) Vox populi. *Nature* 75:450–451.

Gonzalez R, Wu G (1999) On the shape of the probability weighting function. *Cognitive Psych.* 38(1):129–166.

Grgić-Hlača N, Castelluccia C, Gummadi KP (2022) Taking advice from (dis) similar machines: The impact of human-machine similarity on machine-assisted decision-making. *Proc. AAAI Conf. Human Comput. Crowdsourcing*, vol. 10, 74–88.

Griffin D, Brenner L (2004) Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making*, vol. 199, 158–177.

Griffin D, Tversky A (1992) The weighing of evidence and the determinants of confidence. *Cognitive Psych.* 24(3):411–435.

Grimon MP, Mills C (2025) Better together? A field experiment on human-algorithm interaction in child protection. Preprint, submitted February 18, https://arxiv.org/abs/2502.08501.

Grofman B, Owen G, Feld SL (1983) Thirteen theorems in search of the truth. *Theory Decision* 15(3):261–278.

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. *Proc. Internat. Conf. Machine Learn.* (PMLR), 1321–1330.

Hasan E, Duhaime E, Trueblood JS (2024a) Boosting wisdom of the crowd for medical image annotation using training performance and task features. *Cognition Res. Principle Implications* 9(1):31.

Hasan E, Eichbaum Q, Seegmiller AC, Stratton C, Trueblood JS (2024b) Harnessing the wisdom of the confident crowd in medical image decision-making. *Decision* 11(1):127.

Hastie R, Kameda T (2005) The robust beauty of majority rules in group decisions. *Psych. Rev.* 112(2):494.

Holmes WR, O'Daniels P, Trueblood JS (2020) A joint deep neural network and evidence accumulation modeling approach to human decision-making with naturalistic images. *Comput. Brain Behav.* 3:1–12.

Hora SC (2004) Probability judgments for continuous quantities: Linear combinations and calibration. *Management Sci.* 50(5): 597–604.

Kameda T, Tsukasaki T, Hastie R, Berg N (2011) Democracy under uncertainty: The wisdom of crowds and the free-rider problem in group decision making. *Psych. Rev.* 118(1):76.

Koehler DJ, Brenner L, Griffin D (2002) The calibration of expert judgment: Heuristics and biases beyond the laboratory. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 686–715.

Koriat A (2012) The self-consistency model of subjective confidence. *Psych. Rev.* 119(1):80.

Kurvers RH, Herzog SM, Hertwig R, Krause J, Carney PA, Bogart A, Argenziano G, et al. (2016) Boosting medical diagnostics by pooling independent judgments. *Proc. Natl. Acad. Sci. USA* 113(31):8777–8782.

Lee MD, Lee MN (2017) The relationship between crowd majority and accuracy for binary decisions. *Judgment Decision Making* 12(4):328–343.

Lee S, Chu Y, Yoo S, Choi S, Choe S, Koh S, Chung K, et al. (2020) Augmented decision-making for acral lentiginous melanoma detection using deep convolutional neural networks. *J. Eur. Acad. Dermatology Venereology* 34(8):1842–1850.

Lichtenstein S, Fischhoff B, Phillips LD (1977) Calibration of probabilities: The state of the art. *Proc. 5th Res. Conf. Subjective Probability Utility Decision Making* (Springer), 275–324.

Lu T, Zhang Y (2024) 1+ 1 > 2? Information, humans, and machines. *Inform. Systems Res.* 36(1):394–418.

Maier-Hein L, Mersmann S, Kondermann D, Bodenstedt S, Sanchez A, Stock C, Kenngott HG, et al. (2014) Can masses of non-experts

train highly accurate image classifiers? A crowdsourcing approach to instrument segmentation in laparoscopic images. *Proc. 17th Internat. Conf. Medical Image Comput. Comput.-Assisted Intervention* (Springer), 438–445.
Meyen S, Sigg DM, Luxburg U, Franz VH (2021) Group decisions based on confidence weighted majority voting. *Cognition Res. Principles Implication* 6:1–13.
Murphy AH (1972) Scalar and vector partitions of the probability score: Part I. Two-state situation. *J. Appl. Meteorology Climatology* 11(2):273–282.
Murray A, Rhymer J, Sirmon DG (2021) Humans and technology: Forms of conjoined agency in organizations. *Acad. Management Rev.* 46(3):552–571.
Nguyen Q, Valizadegan H, Hauskrecht M (2014) Learning classification models with soft-label information. *J. Amer. Medical Inform. Assoc.* 21(3):501–508.
Nitzan S, Paroush J (1982) Optimal decision rules in uncertain dichotomous choice situations. *Internat. Econom. Rev. (Philadelphia)* 289–297.
Park DK, Kim EJ, Im JP, Lim H, Lim YJ, Byeon JS, Kim KO, et al. (2024) A prospective multicenter randomized controlled trial on artificial intelligence assisted colonoscopy for enhanced polyp detection. *Sci. Rep.* 14(1):25453.
Peterson JC, Battleday RM, Griffiths TL, Russakovsky O (2019) Lee KM, Forsyth D, Pollefeys M, Tang X, Kweon IS, Paragios N, Yang MH, Lazebnik S, eds. Human uncertainty makes classification more robust. *Proc. IEEE/CVF Internat. Conf. Comput. Vision* (IEEE, Piscataway, NJ), 9617–9626.
Pleskac TJ, Busemeyer JR (2010) Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psych. Rev.* 117(3):864.
Press G (2021) Centaur labs gets $15 million to improve data for health care AI. *Forbes* (September 3), https://www.forbes.com/sites/gilpress/2021/09/03/centaur-labs-gets-15-million-to-improve-data-for-healthcare-ai/.
Ranjan R, Gneiting T (2010) Combining probability forecasts. *J. Roy. Statist. Soc. Ser. B Statist. Methodology* 72(1):71–91.
Sorkin RD, Hays CJ, West R (2001) Signal-detection analysis of group decision making. *Psych. Rev.* 108(1):183.
Steyvers M, Tejeda H, Kerrigan G, Smyth P (2022) Bayesian modeling of human–AI complementarity. *Proc. Natl. Acad. Sci. USA* 119(11):e2111547119.
Surowiecki J (2005) *The Wisdom of Crowds* (Anchor, New York).
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, et al. (2015) Going deeper with convolutions. Bischof H, Forsyth D, Schmid C, Sclaroff S, Grauman K, Learned-Miller E, Torralba A, Zisserman A, eds. *Proc. IEEE Conf. Comput. Vision Pattern Recognition* (IEEE, Piscataway, NJ), 1–9.
Trueblood JS, Holmes WR, Seegmiller AC, Douds J, Compton M, Szentirmai E, Woodruff M, et al. (2018) The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognition Res. Principle Implications* 3:1–14.
Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS (2014) Forecast aggregation via recalibration. *Machine Learn.* 95: 261–289.
Tversky A, Fox CR (1995) Weighing risk and uncertainty. *Psych. Rev.* 102(2):269.
Vickers D (2014) *Decision Processes in Visual Perception* (Elsevier, Amsterdam).
Wang G, Kulkarni SR, Poor HV, Osherson DN (2011) Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis (Oxford)* 8(2):128–144.
Winkler RL, Grushka-Cockayne Y, Lichtendahl KC Jr, Jose VRR (2019) Probability forecasts and their combination: A research perspective. *Decision Analysis (Oxford)* 16(4):239–260.
Xu Z, Xia M (2012) Hesitant fuzzy entropy and cross-entropy and their use in multiattribute decision-making. *Internat. J. Intelligent Systems* 27(9):799–822.

Gunnar P. Epping is a data scientist at Centaur.ai and earned his PhD from Indiana University in 2025. His research examines collective intelligence and machine learning, with a focus on how human biases in data annotation propagate to artificial intelligence models in deployment.

Andrew Caplin is Silver professor of economics at New York University and a fellow of the Econometric Society. In addition to this, he is director of the Behavioral Economics Program, National Bureau of Economic Research and the Sloan-Nomis Program on the Cognitive Foundations of Economic Behavior. He has proposed innovative methods of housing finance. The work he is doing to introduce cognitive economics as a unique academic and applied approach is supported by the Alfred P. Sloan Foundation.

Erik Duhaime is co-founder and chief executive officer of Centaur.ai, which supports the development of artificial intelligence by accurately annotating medical and scientific data at scale. The concept for Centaur was developed during his PhD research at the MIT Center for Collective Intelligence. Inspired by his wife's experience in medical training, he ran experiments on how to best combine the opinions of multiple people and AI algorithms for tasks like classifying skin lesions for cancer.

William R. Holmes is an associate professor of cognitive science and mathematics at Indiana University. He is an interdisciplinary applied scientist who uses mathematical and computational tools to study cognitive and biological systems.

Daniel Martin is the Wilcox Family chair in entrepreneurial economics and an associate professor at the University of California, Santa Barbara. He is a behavioral, cognitive, and experimental economist who studies attention and perception (how information is processed) and information disclosure (how information is communicated). His current research explores how human and artificial intelligence interactions are impacted by attention, perception, and information disclosure.

Jennifer S. Trueblood is the Ruth N. Halls professor in the Department of Psychological and Brain Sciences and Cognitive Science Program at Indiana University. Her research combines behavioral experimentation and computational modeling to investigate how individuals make decisions when confronted with multiple complex alternatives. Her work investigates cognitive processes underlying decision making, such as medical image interpretation, consumer behavior, and financial decision making.