Bayesian Modeling of Human-AI Complementarity

Mark Steyvers[a,1], **Heliodoro Tejeda**[a], **Gavin Kerrigan**[b], **and Padhraic Smyth**[b]

[a]Department of Cognitive Sciences, University of California, Irvine; [b]Department of Computer Science, University of California, Irvine

This manuscript was compiled on June 24, 2021

Artificial intelligence (AI) and machine learning models are being increasingly deployed in real-world applications. In many of these applications, there is strong motivation to develop hybrid systems in which humans and AI algorithms can work together, leveraging their complementary strengths and weaknesses. We develop a Bayesian framework for combining the predictions and different types of confidence scores from humans and machines. The framework allows us to investigate the factors that influence complementarity, where a hybrid combination of human and machine predictions leads to better performance than combinations of human or machine predictions alone. We apply this framework to a large-scale data set where humans and a variety of convolutional neural networks perform the same challenging image classification task. We show empirically and theoretically that complementarity can be achieved even if the human and machine classifiers perform at different accuracy levels as long as these accuracy difference falls within a bound determined by the latent correlation between human and machine classifier confidence scores. In addition, we demonstrate that hybrid human-machine performance can be improved by differentiating between the errors that humans and machine classifiers make across different classes. Finally, our results show that eliciting and including human confidence ratings improves hybrid performance in the Bayesian combination model. Our approach is applicable to a wide variety of classification problems involving human and machine algorithms.

Human-AI Complementarity | Bayesian Modeling | Image Classification

There has been significant progress over the past decade in the development of machine learning and artificial intelligence (AI) techniques, particularly those based on deep learning methods (1). This has led to new and more accurate methods for addressing problems in areas such as computer vision (2), speech recognition (3), and natural language processing (4). In turn, these techniques are increasingly embedded in commercial real-world applications, ranging from autonomous driving to customer service chatbots (5, 6). While these approaches have produced impressive gains in testbed performance metrics, such as predictive accuracy, it is broadly acknowledged that these approaches have systematic weaknesses and blind-spots (7–9). For example, state-of-the-art deep learning classifiers for images and text can fail in surprising and unpredictable ways (10–12).

Thus, hybrid systems where AI algorithms and humans work in partnership are gaining prominence as a focus of both AI and human-computer interaction research (13–17), providing opportunities for more human-centered approaches in the overall design of AI systems (18). An emerging theme in this work is the idea that for many problems, ranging from high-risk (medical decisions, autonomous driving) to low-risk (automated recommendations on what product or movie to select next), systems that allow humans and AI algorithms to work together are likely to occupy an important part of the spectrum between full autonomy and no autonomy (19–23).

Indeed, there is empirical evidence to suggest that human and machine algorithms working together can be more effective than either working alone, for tasks as varied as face recognition (24), sports prediction (25), diagnostic imaging (26), and classifying astronomical images (27). This prior work demonstrates that humans and machine algorithms can have complementary strengths and weaknesses, possibly resulting from using different sources of information as well as different strategies to process information. For example, in image classification tasks, the differences in processing strategies by humans and machine classifiers lead to different types of errors made by each, even though their overall level of accuracy is similar (28). As a result, a variety of new ideas have emerged on designing crowdsourcing platforms which can leverage algorithmic predictions given limited human resources (29) as well as new algorithmic and theoretical frameworks that optimize machine predictions in the context of working with humans (30–33).

Previous research in decision-making and machine learning has focused on demonstrating the benefits of combining predictions across individuals or algorithms. For example, statistically combining the predictions from a group of individuals often leads to performance better than any individual in the group, especially when the group is diverse (34–37). Similarly, work on ensemble methods in machine learning has shown that combining classifiers is particularly effective when they are less correlated in their predictions (38–41). While much research on human decision-making and machine learning has contributed to our understanding of separate combinations of human (37, 42) or algorithm predictions (43), less is known about the factors that influence hybrid combinations of both.

To systematically investigate these factors, we develop a

Significance Statement

With the increase in artificial intelligence in real-world applications, there is interest in building hybrid systems that take both human and machine predictions into account. Previous work has shown the benefits of separately combining the predictions of diverse machine classifiers or groups of people. Using a Bayesian modeling framework, we extend these results by systematically investigating the factors that influence the performance of hybrid combinations of human and machine classifiers while taking into account the unique ways human and algorithmic confidence is expressed.

M.S. and P.S. designed research; M.S. and H.T performed research; G.K performed theoretical analysis and M.S., G.K. and P.S wrote the paper.

The authors declare no conflict of interest.
Data deposition: The analysis code and data are available at OSFHome, https://osf.io/2ntrf/?view_only=9ec9cacb806d4a1ea4e2f8acaada8f6c

[1]To whom correspondence should be addressed. E-mail: mark.steyversuci.edu

Bayesian modeling framework that can jointly model human and machine classifier predictions. We apply the framework to a large dataset where humans and a variety of convolutional neural networks (CNNs) perform the same challenging image classification task. CNNs and human visual processing share a number of similarities in terms of their internal representations (44) and the internal representations of CNNs can explain some aspects of human decisions in image classification experiments (45). However, there are also differences in the errors that humans and CNNs make in image classification tasks (28, 46), making image classification an ideal domain to test for complementarity.

With the Bayesian framework, we can empirically and theoretically investigate the conditions that give rise to complementarity. For example, is it better to combine the predictions from a mixture of humans and machine algorithms, leveraging their complementary strengths and weaknesses? Further, when is it better to combine predictions from a group of humans (without algorithms) or a set of machine algorithms (without humans) all performing the same task? Finally, how important is it to differentiate the errors that human and machine algorithms make and how can we combine qualitatively different expressions of confidence across humans and algorithms?

Combining human and machine classifier predictions

The Bayesian combination model we introduce combines the classifications and confidence scores from different ensembles of classifiers, where we use the term "classifier" to refer to either a human or machine classifier. Although this framework can be applied to any number of classifiers, to simplify the analysis we focus on pairs of classifiers: hybrid human-machine (HM) pairs, human-human (HH) pairs, and machine-machine (MM) pairs. For each image, the predictions from the two classifiers in the pair are combined leading to a prediction for the pair.

The modeling approach generates a combined prediction as well as estimates of the latent correlation between classifiers (SI Appendix Fig. S1 provides a schematic overview of the generative process assumed by the model). This correlation captures the dependencies across confidence scores of human and/or machine classifications. For example, if one classifier (human or machine) is confident about the label for a particular image, another classifier (human or machine) might show a similar level of confidence about the label for the same image. The correlation between classifiers is a key characteristic of this latent representation and is estimated for the different pair types (HM, HH, and MM). Previous combination models rely on strong conditional independence assumptions (39, 47) or assume that all predictors have the same output types (43, 48, 49), and hence, fail to address the unique challenges of human-machine combinations. In particular, previous approaches are not applicable when human and machine classifiers provide different types of confidence scores. For example, machine classifiers (including CNNs) typically produce a probability distribution representing the confidence scores across all labels. In contrast, for the human classifiers it is not practical to request confidence scores for all possible labels. Instead, we model a more typical scenario where a human provides a single confidence score associated with the classification. We assume that human confidence is expressed through a small set of ordinal responses (e.g., "low", "medium", "high"), leading to a different type of confidence score compared to the continuous scores provided by the machine classifier. The difference in confidence scoring between human and machine classifiers is modeled by different generative processes for confidence scoring, operating on the same latent representations.

We first consider the problem of combining the predictions from a hybrid human-machine classifier pair. We assume there are N total images and for each image there are L classes. In addition, both the human and classifier are assumed to be noisy labelers relative to the ground truth $z_i \in \{1, ..., L\}$ for each image i. The generative process starts with a bivariate normal model to generate latent logit scores for each label (similar to the logit-normal model (50)). The ground truth z_i determines which of two bivariate normal distributions is used to generate the latent samples λ for instance i. Depending on whether the label matches or mismatches the true label, the bivariate distributions have means $\binom{a_H}{a_M}$ or $\binom{b_H}{b_M}$ respectively:

$$\binom{\lambda_{H,i,j}}{\lambda_{M,i,j}} \sim \begin{cases} \mathcal{N}\left(\binom{a_H}{a_M}, \begin{pmatrix} \sigma_H^2 & \sigma_H\sigma_M\rho_{HM} \\ \sigma_H\sigma_M\rho_{HM} & \sigma_M^2 \end{pmatrix}\right) & \text{if } z_i = j \\ \mathcal{N}\left(\binom{b_H}{b_M}, \begin{pmatrix} \sigma_H^2 & \sigma_H\sigma_M\rho_{HM} \\ \sigma_H\sigma_M\rho_{HM} & \sigma_M^2 \end{pmatrix}\right) & \text{if } z_i \neq j \end{cases} \quad [1]$$

In this generative model, on a per label basis, the logit scores across classifiers are generated from a multivariate normal distribution which captures the dependencies between labels. The covariance matrix captures the dependencies between scores for corresponding labels of the classifiers where ρ is the (latent) correlation between the two classifiers, and σ^2 the variance of logit scores. Across the labels, the logit scores for the label that matches the true label have means a and the logit scores for all other labels have means b. The difference $a - b$ determines the ability of the classifier to discriminate between labels. The logit scores λ are transformed to (normalized) probability confidence scores (i.e., the estimated class probabilities) for both the human and machine classifier:

$$\begin{aligned} \gamma_{H,i,j} &\propto \exp(\lambda_{H,i,j})/(1 + \exp(\lambda_{H,i,j})) \\ \gamma_{M,i,j} &\propto \exp(\lambda_{M,i,j})/(1 + \exp(\lambda_{M,i,j})) \end{aligned} \quad [2]$$

For the machine classifier, the γ_M confidence scores are observable for all labels, as produced by the output of the CNN models. For the human classifier, the γ_H confidence scores are latent and assumed to form the basis for generating a single decision and a confidence rating associated with the decision. To produce the human classification y, we first apply a softmax rule to the latent confidence scores:

$$y_i \sim \text{Categorical}\left(\frac{e^{\gamma_{i,1}/\tau}}{\sum_j^L e^{\gamma_{i,j}/\tau}}, \ldots, \frac{e^{\gamma_{i,L}/\tau}}{\sum_j^L e^{\gamma_{i,j}/\tau}}\right) \quad [3]$$

where we have suppressed the H index for readability. The temperature parameter τ controls the degree to which the label with the highest probability score determines the classification, modeling the noise that arises in a number of human decision-making contexts (45, 51).

To model the human confidence ratings, we use an ordered probit model that probabilistically maps the latent probability score γ_{i,y_i} corresponding to the classification made by the human to an ordinal confidence rating, r_i. For our data, we have three confidence ratings (1="Low", 2="Medium", and 3="High") generated according to:

$$r_i \sim \text{OrderedProbit}(\gamma_{i,y_i}, c, \delta) \quad [4]$$

where the parameters c determine the intervals that map the latent confidence score into a confidence rating and δ determines the sharpness of the rating probability curves, i.e., the degree of randomness in the probabilistic mapping from the confidence score to a rating (see SI Appendix for details).

The preceding description of the model applies to the case of a hybrid HM pair. For MM pairs, the human is replaced by another machine classifier in Equations 1-2 and Equations 3-4 are left unused. For HH pairs, the machine classifier in Equations 1-2 is replaced by another human and Equations 3-4 are applied separately to each individual human classifier.

Theoretical limits of complementarity. While our Bayesian model allows us to combine human and machine predictions, the general conditions under which complementarity arises are not immediately clear. In this section, we analyze our combination model and derive a condition characterizing complementarity in terms of the accuracies and latent correlations of the classifiers.

Specifically, let H_1 and H_2 be two human classifiers and let M_1 and M_2 be two machine classifiers. For any pair of classifiers $C_1, C_2 \in \{H_1, H_2, M_1, M_2\}$, the accuracy of the combined pair of C_1 and C_2 is represented by A_{C_1,C_2}. We have complementarity if for some $H \in \{H_1, H_2\}$ and some $M \in \{M_1, M_2\}$, we have $A_{H,M} > \max\{A_{H_1,H_2}, A_{M_1,M_2}\}$. In our analysis, we assume that H_1 and H_2 are exchangeable, as well as M_1 and M_2. Under additional mild assumptions, we derive a necessary and sufficient condition for complementarity in terms of the individual classifier accuracies and correlations (See SI Appendix for a detailed proof and discussion of our assumptions).

Our main theoretical result is that the accuracy of the Bayesian combination pair for any unique classifiers C_1 and C_2 can be expressed as:

$$A_{C_1,C_2} = \int_{-\infty}^{\infty} \Phi(x)^{L-1} \phi(x - r_{C_1,C_2})\, dx \qquad [5]$$

where $\Phi(\cdot)$ represents the CDF of a standard Gaussian random variable and $\phi(\cdot)$ represents its PDF. The variable r_{C_1,C_2}, which depends on the parameters of our combination model, is defined for each pair type as:

$$r_{H_1,H_2} = \frac{|a_H|}{\sigma_H}\sqrt{\frac{2}{1+\rho_{HH}}} \qquad r_{M_1,M_2} = \frac{|a_M|}{\sigma_M}\sqrt{\frac{2}{1+\rho_{MM}}}$$
$$r_{HM} = \frac{1}{\sigma\sqrt{1-\rho_{HM}}}\sqrt{\frac{a_H^2 + a_M^2 - 2a_H a_M \rho_{HM}}{1+\rho_{HM}}} \qquad [6]$$

Although the integral in Eq. (5) does not have an analytical solution, it can be shown that $A_{C_1,C_2} > A_{C_1',C_2'}$ if and only if $r_{C_1,C_2} > r_{C_1',C_2'}$. Hence, complementarity is equivalent to the condition $r_{H,M} > \max\{r_{H_1,H_2}, r_{M_1,M_2}\}$. In the Supplement, we further analyze the condition $r_{HM} > \max\{r_{HH}, r_{MM}\}$, allowing us to predict complementarity from given model parameters.

Note that by Eq. (6), increasing the non-hybrid correlations (ρ_{MM} and ρ_{HH}) will always cause the non-hybrid pair accuracies to decrease, thus making complementarity easier to achieve. Similarly, increasing r_{HM} will increase the hybrid accuracy. However, since r_{HM} has a more complex dependence on ρ_{HM}, increasing the hybrid correlation ρ_{HM} will cause A_{HM} to decrease if and only if $\min\left(\frac{a_M}{a_H}, \frac{a_H}{a_M}\right) > \rho_{HM}$. Intuitively, the ratios a_M/a_H and a_H/a_M control the relative human-model performance, and higher human-model correlations can be beneficial if the humans and models have vastly different levels of performance.

Results

To empirically verify our theoretical results and to further investigate the factors that influence complementarity, we collected a large data set of human and machine classification decisions for a set of 4800 images. To create variability in machine classifier performance, we selected a number of well-known benchmark convolutional neural network (CNN) architectures (1) for image classification, representative of the recent state-of-the-art in machine classification performance.

To examine conditions for complementarity, we created a number of experimental conditions that lead to variability in performance for human and machine classifiers. One such manipulation is based on adding varying degrees of image noise (46), affecting both human and machine classifier performance. In addition, the classifiers were tuned to the image noise to varying degrees in order to create additional variations in machine classifier performance (SI Appendix, Fig. S4).

Human and machine classifiers make different types of errors. Even at comparable levels of performance, human participants and machine classifiers make different types of errors. Figure 1 shows examples of human-machine algorithm complementarity. The images in 1(a) are challenging for humans but relatively easy for machine classifiers. For all of these images, human accuracy and confidence was low (all six human participants made a low confidence classification and at most one out of six human participant made a correct judgment), but machine accuracy was high (at least four out of five machine classifiers made a correct classification for any of these images). The images in 1(b) are challenging for machine classifiers but relatively easy for humans. All six human judges made a correct and high confidence classification whereas at most one out of five machine classifiers models made a correct classification for each of these images.

Hybrid combinations of human and machine classifiers lead to high accuracy. For the Bayesian combination model, we created a number of data sets based on three different types of pairs: human-human (HH), human-machine (HM), and machine-machine (MM) classifier pairs. To combine the classifications from a pair of human and/or machine classifiers, we use Markov Chain Monte Carlo for inference and obtain samples from the posterior distribution (SI Appendix). Of particular interest is the inferred latent ground truth z label and correlation ρ. Figure 2 shows the out-of-sample accuracy results, based on four-fold cross-validation, of the Bayesian combination model. The results are based on low levels of image noise ($\Omega = 80$) and with CNNs that are fine-tuned for one epoch (see SI Appendix, Figs. S9-11 for the results based on other levels of image noise and fine-tuning).

Our first finding is that the hybrid pairs of human and machine classifiers perform at a high accuracy relative to non-hybrid combinations such as two humans or two machine classifiers, especially for high levels of image noise. However, for CNNs such as Alexnet (SI Appendix), the hybrid combination of Alexnet and a human classifier does not always

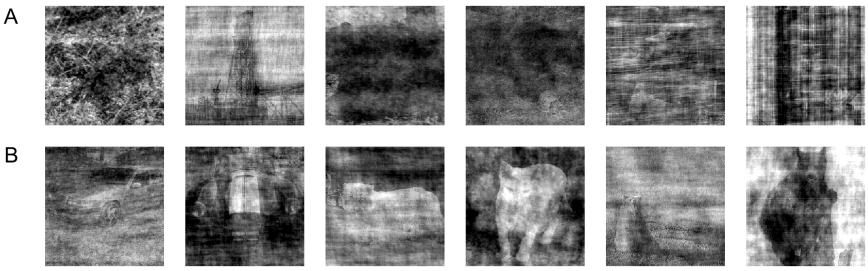

Fig. 1. Examples of human and machine classifier complementarity. (A) examples of images that are challenging for humans, but relatively easy for machine classifiers. Correct answers in reading order are: bird, boat, bear, bear, oven, and oven. (B) examples of images that are challenging for machine classifiers, but relatively easy for humans. Correct answers in reading order are: car, car, cat, cat, bear, and bear. The machine classifiers in both examples were tuned for one epoch on the noisy images.

exceed the performance of a combination of two humans. For this combination, the low baseline performance of the Alexnet classifier does not produce complementarity. The results also show that a combination of two humans leads to better performance than a single human, demonstrating the utility of the human confidence scores — when two human observers differ in confidence, the Bayesian combination model infers that the higher confidence classification is more likely to be correct.

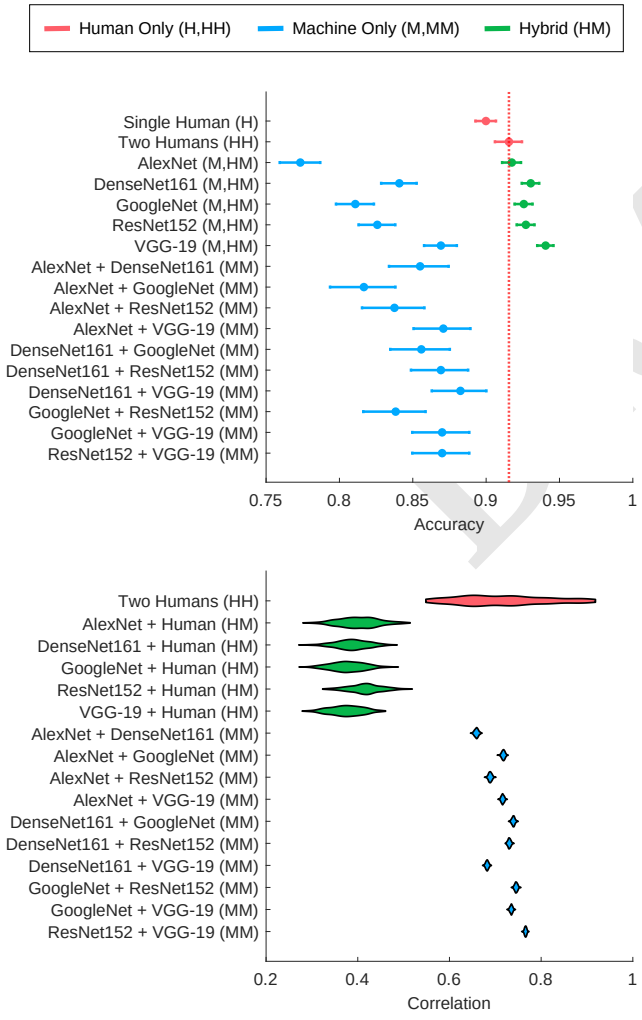

Fig. 2. Accuracy results (top panel) and posterior distributions over correlations (bottom panel) from the Bayesian combination model. Results are broken down by type of classifiers: single human (H), two humans (HH), human-machine classifier (HM), and pairs of machine classifiers (MM). Error bars in the top panel reflect 95% confidence interval of the mean based on a binomial model.

Hybrid combinations of human and machine classifiers lead to low latent correlations. Our second finding is that human and machine classifiers produce lower latent correlations than humans do with each other, or than machine classifiers with each other, demonstrating the utility of combining human and machine predictions – the predictions of hybrid combinations of human-machine classifiers are more independent than the predictions among humans and machine classifiers alone. Figure 2, bottom panel, shows the mean posterior correlations (ρ) between classifier combinations. The hybrid human-machine pairs are correlated less (posterior mean around 0.4) than human-only (posterior mean around 0.7) or machine-only pairs (posterior mean between 0.65 and 0.75). Note that the posterior distributions for the machine classifier correlations are associated with low uncertainty due to the availability of a full set of confidence scores across labels for the machine classifiers. In contrast, for the human classifier, only a single confidence rating is available providing less information to estimate the latent correlational structure.

The inferred pattern of correlation does not critically depend on the representation of the confidence scores. Having only a single continuous confidence score (associated with the classification) and discretizing the machine confidence scores into a small set of ordinal categories, analogous to the human confidence scores does not change the qualitative pattern of results (SI Appendix). This illustrates that the results are robust to different approaches for assessing confidence.

Accuracy difference between classifiers affects complementarity. In our third result, we show empirically how accuracy differences between classifiers lead to complementarity and compare the results with theoretical predictions. Figure 3 shows the observed and predicted complementarity results for a number of hybrid pairs, where the pairs vary in terms of the individual accuracy of the human and machine classifiers composing the pair. Each individual point in the graph is based on the performance of individual classifiers H and M as well as classifier pairs HH, HM, and MM', where M and M' are two different types of CNN classifiers. Complementarity is observed if the hybrid combination HM outperforms the combinations consisting of human or machine classifiers alone: $A_{H,M} > A_{H,H}$ and $A_{H,M} > A_{M,M'}$. To understand how complementarity varies as a function of the difference between human and machine classifier performance, Figure 3 shows the out-of-sample results for 320 comparisons by crossing levels of fine-tuning (4), levels of image noise (4) with CNN pairs (20).

The shaded area in Figure 3 shows that there is a relatively narrow band of performance difference that produces complementarity (see SI Appendix for computational details). The human and machine classifiers need to perform at similar levels in order to produce a hybrid human-machine pair that is more

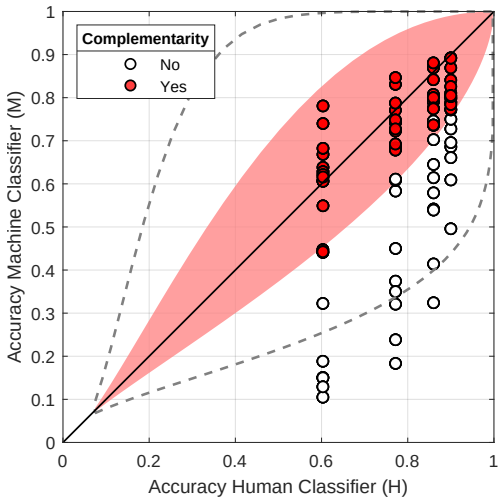

Fig. 3. Observed and predicted complementarity as a function of human and machine classifier accuracy. Circles indicate observed accuracy across different datasets, where filled circles indicate combinations where out-of-sample accuracy of the hybrid human-machine pair outperforms pairs of human-human and machine-machine pairs. The colored area shows the area of complementarity as predicted by theory based on $\rho_{HM} = 0.33$, $\rho_{HH} = 0.62$, and $\rho_{MM'} = 0.71$, approximately matching the correlations inferred by the Bayesian combination model. The dashed line shows the predicted area of complementarity for a best-case situation where the latent human and model predictions are uncorrelated, $\rho_{HM} = 0$, and the non-hybrid correlations remain the same ($\rho_{HH} = 0.62$, $\rho_{MM'} = 0.71$). The diagonal line indicates points of equivalent single human and model performance.

accurate than either two humans or two machine classifiers. These results strongly depend on the correlations between human and machine classifier. For example, in a hypothetical scenario where the human-machine classifier correlation is zero, the zone of complementarity will grow (dotted line). However, note that even in this best-case scenario, there are still limits on the accuracy differences that produce complementarity.

Differentiating between human and machine classifier errors and confidence scores improves hybrid human-machine performance. In our fourth and final finding, we consider how the performance of the hybrid human-machine pairs depends on a number of combinations of different factors: 1) the presence of a class-specific error-model that can correct for human and machine-classifier errors and biases for individual labels, 2) the presence of human confidence scores; and 3) the presence of machine classifier scores. Table 1 shows the out-of-sample accuracy of a hybrid pair when systematically varying these three factors. See SI Appendix for details on models and experimental methodology. The results are averaged over the 5 machine classifiers (Supplement shows results broken down by individual classifiers). Each of the three factors contributes to an improvement in performance of the hybrid ensemble, especially for high noise conditions. In addition, each of these factors has an independent effect on hybrid performance. Table 2 shows the statistical analysis of the relative effects of the three factors on hybrid performance. All three factors are significant. The availability of machine confidence has a larger effect on performance than either the availability of human confidence or an error model. The difference in confidence scoring likely contributes to this difference – the machine classifiers express confidence scores across all labels simultaneously whereas the human participants express only a single confidence score associated with the decision. In addition, the human confidence and error model contribute about the same in performance to hybrid performance. Thus, a simple way to boost performance of hybrid human-machine classifiers is to elicit human confidence ratings.

Table 1. Accuracy for human-machine classifier combinations across image noise and different types of combination models that vary the presence or absence of an error model, human confidence scores, and machine classifier confidence scores. The results are averaged across the 5 machine classifiers. Each accuracy result is based on 36,000 observations.

Error Model	Human Confidence	Machine Confidence	Image Noise (ω) 80	95	110	125
✓	✓	✓	0.933	0.906	0.850	0.748
✗	✓	✓	0.927	0.899	0.841	0.722
✓	✗	✓	0.928	0.902	0.844	0.738
✗	✗	✓	0.925	0.895	0.830	0.707
✓	✓	✗	0.911	0.883	0.823	0.701
✗	✓	✗	0.903	0.876	0.815	0.686
✓	✗	✗	0.901	0.872	0.805	0.674
✗	✗	✗	0.895	0.858	0.769	0.636

Discussion

Previous work has shown the benefits of separately combining the predictions of diverse machine classifiers (38–41) or groups of people (34–37). In this work, we extend these results by systematically investigating the factors that influence the performance of hybrid combinations of machine and human classifiers. We collected a large-scale behavioral and machine classifier data set where both humans and machine classifiers make predictions for the same data. The results showed that even if performance from a human exceeds the performance of a machine classifier, adding the predictions from the machine classifier to a single human can lead to better performance than combining the predictions of two humans. The converse is also true. Even if a machine classifier outperforms humans, a hybrid human-machine pair can still outperform the predictions from a combination of machine classifiers that are all individually outperforming a single human.

Our results have implications for algorithmic systems that have not yet achieved human-level accuracy (e.g. (52)). Starting with a human predictor, adding algorithmic predictions (that are less accurate than the human) may be more beneficial than adding additional human predictors. Thus, the

Table 2. Effect of including class-specific error model, human confidence, and machine classifier confidence scores on hybrid human-machine performance.

Predictors	Accuracy (Log Odds) Estimate	CI	P
Intercept	1.368	1.358 to 1.377	$< .001$
Error Model	0.101	0.091 to 0.110	$< .001$
Human Confidence	0.104	0.094 to 0.114	$< .001$
Machine Confidence	0.257	0.247 to 0.266	$< .001$
Observations	1,152,000		

benchmark for evaluating AI algorithms need not necessarily be human-level performance. If an algorithm does not achieve human-level accuracy it can still lead to increased accuracy in combined hybrid predictions. Conversely, our results also indicate that once AI approaches have exceeded human performance in particular domains, this does not imply that human judgment is no longer useful in hybrid human-machine systems.

However, there are limits to the scope of complementarity. Prior work has shown empirically that hybrid human-machine algorithm systems do not always lead to superior performance (33, 53, 54). Our results in this paper go beyond these earlier studies, both theoretically and empirically, and show specifically what factors contribute to complementarity (25). In particular, the key limiting factor for complementarity is the degree of correlation between human and machine classifier predictions. A large correlation leads to limits on the accuracy difference between classifiers that can support complementarity. This result has implications for human-AI collaborative settings where algorithms are used as decision aids (54, 55). Effective AI advice should not only be accurate but also be as independent as possible from human judgment. Independence of the AI component from the human could for example be increased by leveraging different mechanisms to produce predictions or changing the objective function for the AI model (31). Interestingly, the goal of decreasing the correlation between human and algorithmic predictions stands in contrast with modeling natural intelligence, where the goal is to create computational models that mimic human internal processing mechanisms (28).

Another important factor is the role of both human and machine classifier confidence scores. While machine classifier scores have been used before in hybrid human-machine systems (56), human confidence is often not elicited (29, 57). However, our results show that human confidence ratings can significantly increase hybrid performance and is as effective in improving combined performance as inferring an explicit error model that can correct for class-specific errors and biases. Confidence scores allow differing abilities of human and machine classifiers to be resolved at the level of individual instances.

Overall, our results add to a growing literature showing the advantages of combining human and AI predictions in areas such as crowdsourcing (29, 57, 58), providing a framework for assessing hybrid combinations of human and machine predictions, with potential applications in high-stakes domains such as medicine (59–61) and the justice system (62, 63).

Materials and Methods

The original and preprocessed versions of the data can be accessed at: https://osf.io/2ntrf/?view_only=9ec9cacb806d4a1ea4e2f8acaada8f6c

Images for Experiments. There are 1200 unique images total in our dataset, divided equally into 16 classes (chair, oven, knife, bottle, keyboard, clock, boat, bicycle, airplane, truck, car, elephant, bear, dog, cat, and bird). The images and categories are based on a subset of the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 database (64). As ground truth labels we used the original labels from the ILSRVR database. To create a more challenging classification task for both the human participants and machine classifiers, images were distorted by phase noise at each spatial frequency, where the phase noise is uniformly distributed in the interval $[-\omega, \omega]$ (65). Four levels of phase noise, $\omega = \{80, 95, 110, 125\}$, were applied to each of the 1200 unique images, resulting in 4800 images (see SI Appendix, Fig. S3 for examples).

Behavioral Image Classification Experiment. The behavioral image classification dataset consists of 28,997 human classifications from a total of 145 participants. Each participant classified 200 images into the 16 categories. For each classification, participants also provided a discrete confidence level ("low", "medium","high"). The behavioral classification dataset contains at least six human classifications for each of the 4800 images. Human performance decreases as a function of image noise and accuracy varies systematically as a function of expressed confidence (Supplementary Results), showing that confidence is related to decisional uncertainty.

Machine Classifier Predictions. We created a set of machine classifier predictions for the 4800 images and the set of 16 classes in the behavioral dataset. For each image, each classifier produces a probability vector over the 16 classes, containing the confidence scores for each class. The class associated with the highest probability corresponds to the classification for the image. To vary performance of the machine classifiers relative to human performance, we selected five different machine classifiers pre-trained for ImageNet: AlexNet (66), DenseNet161 (67), GoogleNet (68), ResNet152 (69), and VGG-19 (70). To create additional levels of performance variation, we retrained the models to varying degrees to adapt to the image distortions. For each of the five classifiers, we retrained four variants of each model, based on how many passes through the noisy images data (epochs) are used during stochastic gradient training, producing in effect four variants that are adapted/fine-tuned to varying degrees of noise. The models were fine-tuned for either 0 epochs (baseline), between 0 and 1 epochs, 1 epoch, and 10 epochs. The second level of finetuning (0-1 epochs) was based on a checkpoint during training before 1 epoch was reached, leading to a performance level intermediate between baseline and 1 epoch of training. The different machine classifiers produce a variety of performance levels relative to human performance, with some finetuned VGG-19 and DenseNet161 classifiers exceeding human performance at the high image distortion levels (SI Appendix).

1. Y LeCun, Y Bengio, G Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
2. J Malik, Technical perspective: What led computer vision to deep learning? *Commun. ACM* **60**, 82–83 (2017).
3. L Deng, G Hinton, B Kingsbury, New types of deep neural network learning for speech recognition and related applications: An overview in *2013 IEEE international conference on acoustics, speech and signal processing*. (IEEE), pp. 8599–8603 (2013).
4. J Hirschberg, CD Manning, Advances in natural language processing. *Science* **349**, 261–266 (2015).
5. RG Smith, J Eckroth, Building ai applications: Yesterday, today, and tomorrow. *AI Mag.* **38**, 6–22 (2017).
6. E Brynjolfsson, T Mitchell, What can machine learning do? workforce implications. *Science* **358**, 1530–1534 (2017).
7. N Papernot, et al., The limitations of deep learning in adversarial settings in *2016 IEEE European symposium on security and privacy (EuroS&P)*. (IEEE), pp. 372–387 (2016).
8. T Serre, Deep learning: the good, the bad, and the ugly. *Annu. review vision science* **5**, 399–426 (2019).
9. WE Zhang, QZ Sheng, A Alhazmi, C Li, Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intell. Syst. Technol. (TIST)* **11**, 1–41 (2020).
10. B Recht, R Roelofs, L Schmidt, V Shankar, Do imagenet classifiers generalize to imagenet? in *International Conference on Machine Learning*. (PMLR), pp. 5389–5400 (2019).
11. D Hendrycks, K Zhao, S Basart, J Steinhardt, D Song, Natural adversarial examples in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2021)*. (2021).
12. MT Ribeiro, T Wu, C Guestrin, S Singh, Beyond accuracy: Behavioral testing of nlp models with checklist in *Proceedings of the ACL Conference*. (2021).
13. MO Riedl, Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Technol.* **1**, 33–36 (2019).
14. G Bansal, et al., Beyond accuracy: The role of mental models in human-ai team performance in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7, pp. 2–11 (2019).
15. G Lee, C Mavrogiannis, SS Srinivasa, Towards effective human-ai teams: The case of collaborative packing. *arXiv preprint arXiv:1909.06527* (2019).
16. R Zhang, NJ McNeese, G Freeman, G Musick, "an ideal human" expectations of ai teammates in human-ai teaming. *Proc. ACM on Human-Computer Interact.* **4**, 1–25 (2021).
17. Z Zahedi, S Kambhampati, Human-ai symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990* (2021).
18. B Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Human–Computer Interact.* **36**, 495–504 (2020).
19. E Kamar, Directions in hybrid intelligence: Complementing ai systems with human intelligence. in *Proceedings of IJCAI*. pp. 4070–4073 (2016).
20. I Rahwan, et al., Machine behaviour. *Nature* **568**, 477–486 (2019).
21. M Johnson, A Vera, No ai is an island: the case for teaming intelligence. *AI Mag.* **40**, 16–28 (2019).

22. T O'Neill, N McNeese, A Barron, B Schelble, Human–autonomy teaming: A review and analysis of the empirical literature. *Hum. Factors*, 0018720820960865 (2020).
23. M De-Arteaga, R Fogliato, A Chouldechova, 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*. pp. 1–12 (2020).
24. PJ Phillips, et al., Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci.* **115**, 6171–6176 (2018).
25. Y Nagar, TW Malone, Making business predictions by combining human and machine intelligence in prediction markets in *Thirty Second International Conference on Information Systems*. (Association for Information Systems), (2011).
26. BN Patel, et al., Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit. Medicine* **2**, 1–10 (2019).
27. DE Wright, et al., A transient search using combined human and machine classifications. *Mon. Notices Royal Astron. Soc.* **472**, 1315–1323 (2017).
28. R Geirhos, K Meding, FA Wichmann, Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Adv. Neural Inf. Process. Syst.* **33** (2020).
29. L Trouille, CJ Lintott, LF Fortson, Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proc. Natl. Acad. Sci.* **116**, 1902–1909 (2019).
30. B Wilder, E Horvitz, E Kamar, Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).
31. G Bansal, B Nushi, E Kamar, E Horvitz, DS Weld, Optimizing ai for teamwork. *arXiv preprint arXiv:2004.13102* (2020).
32. A De, P Koley, N Ganguly, M Gomez-Rodriguez, Regression under human assistance. in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*. pp. 2611–2620 (2020).
33. G Bansal, et al., Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779* (2020).
34. I Aggarwal, AW Woolley, CF Chabris, TW Malone, The impact of cognitive style diversity on implicit learning in teams. *Front. Psychol.* **10**, 112 (2019).
35. L Hong, SE Page, Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci.* **101**, 16385–16389 (2004).
36. P Lamberson, SE Page, Optimal forecasting groups. *Manag. Sci.* **58**, 805–810 (2012).
37. CP Davis-Stober, DV Budescu, SB Broomell, J Dana, The composition of optimally wise crowds. *Decis. Analysis* **12**, 130–143 (2015).
38. K Tumer, J Ghosh, Error correlation and error reduction in ensemble classifiers. *Connect. Sci.* **8**, 385–404 (1996).
39. J Kittler, M Hatef, RP Duin, J Matas, On combining classifiers. *IEEE Transactions on Pattern Analysis Mach. Intell.* **20**, 226–239 (1998).
40. G Brown, JL Wyatt, P Tino, Y Bengio, Managing diversity in regression ensembles. *J. Mach. Learn. Res.* **6** (2005).
41. Y Ren, L Zhang, PN Suganthan, Ensemble classification and regression-recent developments, applications and future directions. *IEEE Comput. intelligence magazine* **11**, 41–53 (2016).
42. BM Turner, M Steyvers, EC Merkle, DV Budescu, TS Wallsten, Forecast aggregation via recalibration. *Mach. Learn.* **95**, 261–289 (2014).
43. HC Kim, Z Ghahramani, Bayesian classifier combination in *Artificial Intelligence and Statistics*. pp. 619–627 (2012).
44. N Kriegeskorte, Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
45. RM Battleday, JC Peterson, TL Griffiths, Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat. Commun.* **11**, 1–14 (2020).
46. R Geirhos, et al., Generalisation in humans and deep neural networks in *Advances in Neural Information Processing Systems*. pp. 7538–7550 (2018).
47. Z Oravecz, J Vandekerckhove, WH Batchelder, Bayesian cultural consensus theory. *Field Methods* **26**, 207–222 (2014).
48. O Sagi, L Rokach, Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, e1249 (2018).
49. KM Ting, IH Witten, Issues in stacked generalization. *J. artificial intelligence research* **10**, 271–289 (1999).
50. J Atchison, SM Shen, Logistic-normal distributions: Some properties and uses. *Biometrika* **67**, 261–272 (1980).
51. ND Daw, JP O'doherty, P Dayan, B Seymour, RJ Dolan, Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
52. BE Bejnordi, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
53. S Tan, J Adebayo, K Inkpen, E Kamar, Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).
54. Y Zhang, QV Liao, RK Bellamy, 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*. pp. 295–305 (2020).
55. V Lai, C Tan, On human predictions with explanations and predictions of machine learning models: A case study on deception detection in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. pp. 29–38 (2019).
56. D Madras, T Pitassi, RS Zemel, Predict responsibly: Improving fairness and accuracy by learning to defer in *NeurIPS*. (2018).
57. M Willi, et al., Identifying animal species in camera trap images using deep learning and citizen science. *Methods Ecol. Evol.* **10**, 80–91 (2019).
58. JW Vaughan, Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.* **18**, 7026–7071 (2017).
59. J Wilkinson, et al., Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digit. Heal.* (2020).
60. E Beede, et al., A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy in *Proceedings of the 2020 CHI conference on human factors in computing systems*. pp. 1–12 (2020).
61. M Nagendran, et al., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368** (2020).
62. Y Hayashi, K Wakabayashi, Can ai become reliable source to support human decision making in a court scene? in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. pp. 195–198 (2017).
63. J Kleinberg, H Lakkaraju, J Leskovec, J Ludwig, S Mullainathan, Human decisions and machine predictions. *The Q. J. Econ.* **133**, 237–293 (2018).
64. O Russakovsky, et al., Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
65. R Geirhos, et al., Generalisation in humans and deep neural networks in *Thirty-second Annual Conference on Neural Information Processing Systems (NeurIPS 2018)*. (Curran), pp. 7549–7561 (2019).
66. A Krizhevsky, I Sutskever, GE Hinton, Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
67. G Huang, Z Liu, L Van Der Maaten, KQ Weinberger, Densely connected convolutional networks in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4700–4708 (2017).
68. C Szegedy, et al., Going deeper with convolutions in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1–9 (2015).
69. K He, X Zhang, S Ren, J Sun, Deep residual learning for image recognition in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016).
70. K Simonyan, A Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).