Trucks Don't Mean Trump: Diagnosing Human Error in Image Analysis

J.D. Zamfirescu-Pereira
University of California, Berkeley
Berkeley, USA

Jerry Chen
Stanford University
Stanford, USA

Emily Wen
Stanford University
Stanford, USA

Allison Koenecke
Microsoft Research and Cornell University
Cambridge, USA

Nikhil Garg
Cornell Tech
New York City, USA

Emma Pierson
Cornell Tech
New York City, USA

ABSTRACT

Algorithms provide powerful tools for detecting and dissecting human bias and error. Here, we develop machine learning methods to to analyze how humans err in a particular high-stakes task: image interpretation. We leverage a unique dataset of 16,135,392 human predictions of whether a neighborhood voted for Donald Trump or Joe Biden in the 2020 US election, based on a Google Street View image. We show that by training a machine learning estimator of the Bayes optimal decision for each image, we can provide an actionable decomposition of human error into bias, variance, and noise terms, and further identify specific features (like pickup trucks) which lead humans astray. Our methods can be applied to ensure that human-in-the-loop decision-making is accurate and fair and are also applicable to black-box algorithmic systems.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Machine learning approaches.*

KEYWORDS

image analysis, human error, diagnosing bias

ACM Reference Format:
J.D. Zamfirescu-Pereira, Jerry Chen, Emily Wen, Allison Koenecke, Nikhil Garg, and Emma Pierson. 2022. Trucks Don't Mean Trump: Diagnosing Human Error in Image Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3531146.3533145

1 INTRODUCTION

Recent work in algorithmic fairness has highlighted many of the ways in which algorithms can be biased and error-prone [16, 18, 20, 51]. However, algorithms also provide powerful tools for detecting and dissecting similarly numerous human errors. Understanding patterns of error in human judgment is of interest to a wide range of fields including psychology, computer science, and behavioral economics [15, 22, 23, 44, 54]. Algorithmic and statistical approaches have uncovered systematic human biases—e.g., race or gender biases—in settings including criminal justice [30, 36, 40, 44, 55, 57], medicine [44, 54, 56], and cultural stereotyping [29]. Previous work has also shown the importance of *variance* across decision-makers, in which different decision-makers make inconsistent judgments about similar tasks [38, 40]. The use of algorithms to diagnose sources of human error – whether systematic bias, variance across humans, or unavoidable noise[1] – has also received increasing attention in the algorithmic fairness community [1, 42], in part because humans and algorithms often work together to make decisions, and so understanding the imperfections of the human in the loop is necessary to achieve overall fairness [19, 31, 34, 45, 64].

Here, we develop algorithmic methods to dissect human error in a particular high-stakes task: image interpretation. Humans frequently make important decisions on the basis of images. Clinicians assess x-rays, MRIs, and other image modalities for signs of disease; drivers and pilots respond to fast-changing visual data; online moderators judge whether images are offensive. Understanding patterns of human error in image interpretation has a wide range of applications, including improving training for decision-makers, building algorithmic decision-aids, and deciding when to ask for a second opinion [12, 58]. However, understanding why and how people err in interpreting images is uniquely challenging. Even defining the salient features in a complex image in an interpretable way [35, 66] is difficult; so is determining how those features influence 1) human decisions and 2) the optimal decisions, and comparing the two in a principled way. Furthermore, in many datasets the ground truth itself is defined based on human judgments (e.g., in radiology tasks, it is often the consensus opinion of radiologists [10]), and so measuring how humans deviate from ground truth is circular.

We draw on a unique new dataset of human judgments about images to develop a method for understanding human error. In March 2021, *The New York Times* ran a quiz asking respondents to predict whether a Google Street View image came from a neighborhood in which a majority voted for Donald Trump or Joe Biden

[1] Throughout this paper, we use "bias" to refer to the case in which humans, on average, misweight features (e.g., race or gender) in making a decision; "variance" to refer to inconsistency across human decision-makers when making the same decision; and "noise" to refer to irreducible errors made even by Bayes optimal decision-makers. "Error" is used as an umbrella term that encompasses all human mistakes, while "avoidable" error refers to just bias and variance.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533145

in the 2020 US election [8]. The resulting dataset includes 10,000 neighborhood images, the ground truth for each image (i.e., the true Trump-Biden vote share), and 16,135,392 anonymized individual human predictions of whether a neighborhood voted for Trump or Biden (with more than a thousand human predictions for each image). This dataset is a rich test-bed for methods development because it has 1) an enormous number of human judgments on many unique images; 2) a reasonable prior likelihood that humans perform suboptimally due to stereotypes (e.g., some respondents told journalists that they viewed American flags as predictive of Trump support, but neighborhoods with prominent American flags were actually split evenly between Biden and Trump [7]); and 3) ground truth labels which are derived independently of human judgment (i.e., based on the actual election results), eliminating circularity concerns. While our dataset represents an ideal setting for validating methods, the methods we develop apply to diagnosing human error in image interpretation more generally, as we discuss, as well as to the related task of diagnosing human error in interpreting tabular data and other non-image data modalities.

Using this dataset, we develop a machine learning method to identify when human decision-makers deviate from the Bayes optimal judgment: that is, when they predict "Trump" even though the probability that a neighborhood voted for Biden based on the image is over 50% (or vice versa).

This task is not equivalent to identifying *ex post* errors, where human predictions simply disagree with the ground truth. For example, suppose a Street View image happened to capture the only Trump-supporting household (with a prominent Trump flag) in a strongly Biden-leaning neighborhood. Then, answering "Trump" would be Bayes optimal even though it would disagree with the ground truth. This distinction is key in identifying potentially fixable human mistakes and the image features inducing those mistakes, as opposed to cases where the image is uninformative. We make the following contributions:

- We propose a method for comparing human decision-making to Bayes optimal decision-making, by first training a machine learning algorithm to estimate the Bayes optimal model and then comparing human decisions to those implied by the estimated Bayes optimal model. We show that even if our estimate of the Bayes optimal model is imperfect, our approach can still provide useful insights into human error as long as our machine learning model adds signal beyond human judgment, a property we verify.
- We use our method to provide an actionable decomposition of human error into bias, variance, and noise terms by extending a classic decomposition of machine learning model error [27]. On the Trump-Biden prediction task, we find that noise and variance are larger contributors to human error than is bias.
- We provide both qualitative and quantitative methods for identifying specific image features which contribute to human error—for example, pickup trucks leading humans to guess "Trump" more than is optimal.
- We analyze, and assess the downsides of, two alternate approaches to diagnosing human error—1) training one model to predict human judgment, training a second model to predict ground truth, and examining deviations between the two models and 2) training a single model to predict the difference between human judgment and ground truth.

While we focus on *human* decision-making in this paper, we note that there is little conceptual difference between assessing, from data, human errors and those of black-box technical systems where only the system's input and output is known. Our method only makes use of the input image, humans' binary judgements, and ground truth—and so could be used, for example, to audit black-box third-party vision APIs which only output binary decisions. As such, our work contributes to a long line of work using algorithmic approaches to audit other computational systems [1, 2, 5, 11, 13, 14, 17, 21, 43, 46, 50, 59, 61].

2 RELATED WORK

Measuring and describing error—and in particular, bias and variance—in human decision-making has been a problem of interest to social scientists for decades, ranging from theoretical models of racial, gender, and other types of discrimination [9] to cognitive heuristics which are employed under uncertainty [63]. Kahneman [37] provides a recent review of common biases in human decision-making, and Kahneman et al. [38] addresses the importance of variance, in which humans make inconsistent judgments about similar problems. Our work is motivated by, and builds on, this literature by using modern machine learning approaches to decompose human error into bias, variance, and noise terms, and then to explain image-specific causes. Below, we summarize the algorithmic communities closest to our work.

Algorithmic approaches to measuring error in human decision-making. In recent years, there has been increasing interest in the algorithmic fairness community in using computational and algorithmic approaches to measure error in *human* decision-making, as part of a broader recognition that such approaches can serve a useful *diagnostic* function in precisely understanding and measuring social problems [1, 29]. Several observations motivate this interest. First, much prior work has argued that algorithmic decision-making is (theoretically if often not practically) more transparent than human decision-making, allowing algorithms to serve as "discrimination detectors" [41, 42, 53]. Second, algorithmic tools are often designed to be used by humans, so understanding the imperfections of the human in the loop is necessary to ensure the system as a whole is fair [19, 45, 64]. Failing to account for human biases can produce algorithms which perform well on retrospective data but yield unexpected or pernicious effects when they are actually used by human decision-makers [60].

Motivated by these observations, there have been numerous examples of using algorithmic approaches to uncover systematic human biases. Algorithms have been used to diagnose broader cultural biases and stereotypes, often through use of word embeddings [11, 14, 25, 29]; this work differs from ours because it studies broader cultural trends but not specific human errors in decision-making. Closer to our own work is the use of algorithms to study human errors in decision-making settings. For example, in criminal justice [30, 36, 40, 44, 55, 57], algorithms have been used to diagnose human error in bail decisions [40, 44] and stop-and-frisk [30, 36],

among other settings. In medicine, algorithms have been used to identify human errors in diagnosing pain [56], diagnosing heart attacks [54], and prescribing asthma treatments [44]. Algorithms have also been used to diagnose and dissect human error in games like chess [6, 48, 49].

Our work builds on this literature by developing and evaluating machine learning methods for diagnosing how humans err in analyzing *image data specifically*; in contrast, the work above focuses on tabular data. As discussed above, analyzing human decisions made from image data poses unique challenges. As a representative illustration of how methods for tabular data do not easily transfer to image data, consider the work of Jung et al. [36], who convincingly demonstrate bias in police search decisions in New York City using tabular data. Based on historical prior knowledge, they specifically assess racial bias, which they quantify by measuring the racial disparities in searches which remain when controlling for (a model's estimate of) a pedestrian's objective risk of carrying contraband. Their approach of assessing human decisions relative to objective risk is conceptually similar to ours (see Mullainathan and Obermeyer [54] for another example of this approach). However, their method cannot be directly applied to our setting because it relies on tabular data with 1) clearly interpretable features and 2) an *a priori* understanding of which features are likely to contribute to human biases (in their setting, race); in contrast, with image data, we have neither of these things. While we confront the unique challenge of image data, many of our methods and observations also apply to tabular data because it is an easier task.

Algorithmic descriptions of human decision-making. Beyond specifically diagnosing human error, algorithmic approaches have also been used to describe human decision-making more broadly [4, 34]. Our approach of analyzing human behavior via a residual with a machine learning model is somewhat similar to that of Agrawal et al. [4]; however, while they use a neural network to smooth high-dimensional noise in empirical human decisions when *predicting human decisions*, we use one as a proxy for the Bayesian optimal decision when *analyzing human error*.

Human-algorithmic collaborations. There is substantial work on creating algorithms which can learn to *complement* humans [24, 31, 34, 62], for example, by learning to defer to a human expert when the human will achieve better performance [47, 52, 65] or by providing automated assessments for human experts to consider in making decisions [32]. Such approaches often implicitly model human error. Our work differs in that it focuses on explicitly describing human error, not in creating algorithms which implicitly model human error while learning to complement humans.

Computer vision interpretability. We describe human and optimal decision-making from images using convolutional image models, and consequently rely on methods for interpreting these models. There is a wide literature on such methods: see Zhang and Zhu [68] for a review. In our primary results, we make use of occlusion mapping [67], a commonly used technique which identifies regions of the image which influence a model's prediction by determining how much the prediction changes in response to masking out regions of the image.

3 PROBLEM SETUP

3.1 Data

The New York Times quiz dataset. Our main dataset consists of the 10,000 neighborhood images available in the *New York Times* quiz, which we partition into 5,000 training images; 1,500 validation images; 1,500 images which we use as a preliminary test set while conducting experiments for this paper; and a holdout test set of 2,000 images which we use only to generate the final results for this paper to minimize overfitting [40, 56]. All our main results are reported on this holdout test set. Each image is retrieved from Google Street View as a composite of four individual views for the same location at viewing angles of 0°, 90°, 180°, and 270°, such that the stitched composite image approximates what human respondents could see in the *New York Times* quiz. Each image is linked to the ground truth Trump-Biden vote share in that neighborhood.[2] Additionally, each neighborhood image has an average of 1,614 corresponding human predictions of whether a neighborhood voted for Trump or Biden.[3]

Augmenting the quiz dataset. To increase the amount of data we have to train our models, we collect an additional external dataset of 52,025 Google Street View images linked to election results (but no human judgments), using a sampling method similar to that of *The New York Times*. See Appendix A.1 for details.

3.2 Notation

Throughout, i indexes neighborhood images and j indexes human judgments about each image. For each image X_i, we have human judgments of whether the majority vote in the associated neighborhood was for Trump or Biden, where $h_{ij} \in \{0, 1\}$ denotes individual human judgments; a 0 indicates Biden, and 1 indicates Trump. We use $\overline{h_i} \in [0, 1]$ to denote the mean human judgment for each image (i.e., the fraction of people who indicated Trump for that image). We also have binary ground truth, $y_i \in \{0, 1\}$, indicating the true majority vote for the corresponding neighborhood; it is also useful to refer to the continuous vote share difference (the fractional Trump vote share minus the fractional Biden share), denoted $y_i^{(\text{cont})} \in [-1, 1]$, with $y_i = \mathbb{1}[y_i^{(\text{cont})} > 0]$. We report additional statistics for ground truth and human judgment in Table 1.

Figure 1 plots $y_i^{(\text{cont})}$ against $\overline{h_i}$, showing that, while human judgment is correlated with true election outcomes, humans are far from omniscient and frequently disagree on each image. However, deviations from $y_i^{(\text{cont})}$ are not enough to conclude that humans are making *avoidable* mistakes – it could be that the images are

[2] We use "neighborhood" throughout to refer to *electoral precinct*, the most granular area in the United States for which election results are publicly available. When we say that a neighborhood voted for a candidate, we mean that a majority of voters in that neighborhood – who voted for either Trump or Biden – voted for that candidate.

[3] While we are able to identify which predictions came from the same *New York Times*-identified human—who each on average made predictions on 10.33 images in our training data—we do not know the order in which the respondent made their predictions on different images. As such, we cannot assess the effect of feedback received by each respondent over their sequence of predictions. Furthermore, *The New York Times* did not collect any additional information on respondents, such as demographic information, location, or IP address. Thus, we cannot assess how performance or human error varies by such covariates.

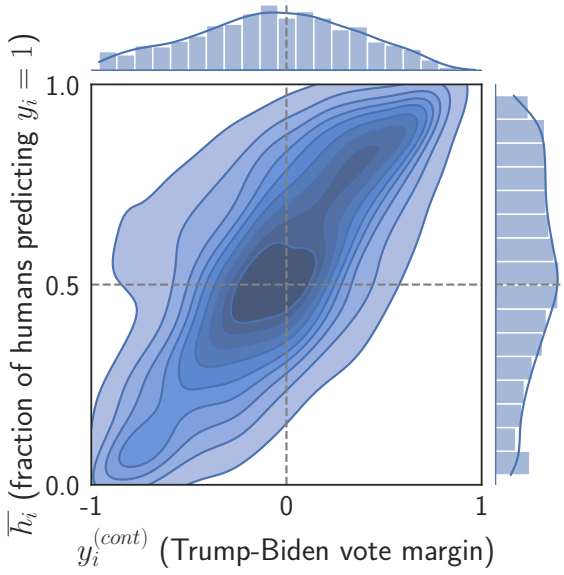


Figure 1: Joint histogram of $y_i^{(\text{cont})}$, the true Trump-Biden vote margin (x-axis), against $\overline{h_i}$, the fraction of humans who guessed "Trump" (y-axis). Humans deviate substantially from both omniscience (which would imply a threshold function at $y_i^{(\text{cont})} = 0$) and perfect agreement with each other (which would imply $\overline{h_i} \in \{0, 1\}$). However, note that this plot is not enough to conclude that humans in aggregate are making avoidable mistakes – it could be that the images are uninformative, and so the errors are due to noise that even a Bayes optimal decision-maker would make. Our methods are designed to separate such noise from avoidable human errors.

uninformative about election outcomes, and so the errors are unavoidable ones that even a Bayes optimal decision-maker would make.

To model such a Bayes optimal decision-maker, we let $p(y_i = 1|X_i)$ denote the probability that the neighborhood corresponding to a certain image voted for Trump; i.e., how often images that look like X_i correspond to a neighborhood where the majority voted for Trump. For example, suppose 80% of neighborhoods with pickup trucks voted for Trump; then, $p(y_i = 1|X_i = \textit{Neighborhood with pickup truck}) = 0.8$.

3.3 Task

A Bayes optimal decision-maker constrained to make binary judgements about each image should predict $y_i = 1$ if and only if $p(y_i = 1|X_i) > 0.5$. Here, we aim to quantify to what extent human decision-makers deviate from Bayes optimality and identify what features of the images lead to these errors. This is a task of primary interest because it quantifies to what extent, and why, humans are making *avoidable* (and potentially fixable) errors in interpreting images. However, this task is challenging for several reasons: 1) the Bayes optimal decision, and in particular the probability $p(y_i = 1|X_i)$, is not directly observable in the data; 2) we also do not observe individual human estimates of $p(y_i = 1|X_i)$, just their binary decisions; and 3) defining salient features in the images is difficult.

Our methods, described next, are designed around these challenges. We note that while our notation and descriptions are particular to our dataset, these characteristics are common for many settings in which humans make decisions using images.

4 METHOD

Our main approach to characterizing human error examines how humans deviate from the Bayes optimal decisions implied by $p(y_i = 1|X_i)$. In Section 4.1 we describe how we fit a model to estimate $p(y_i = 1|X_i)$ and provide evidence of the quality of the estimate. In Section 4.2, we show how our estimate of $p(y_i = 1|X_i)$ can be used to decompose human error into bias, variance, and noise terms. Finally, in Section 4.3, we show how our estimate of $p(y_i = 1|X_i)$ can be used to identify specific image features which lead humans astray.

4.1 Estimating and plotting $p(y_i = 1|X_i)$

Training a model to estimate $p(y_i = 1|X_i)$. We estimate this model in two steps. First, we train a deep learning model to estimate $p(y_i = 1|X_i)$ on the large external dataset collected as described in Section 3.1 (see Appendix A.2 for model training details; we verify that training a model on the large external dataset yields slightly superior performance to training only on the smaller *New York Times* dataset). We use $\hat{f}^{(\text{ext})}(X_i)$ to denote our deep learning model's estimate of $p(y_i = 1|X_i)$. Second, using the *New York Times* training and validation datasets, we fit a simple logistic regression to estimate $p(y_i = 1|X_i)$ using both the deep learning model's prediction $\hat{f}^{(\text{ext})}(X_i)$ and the aggregate human judgment $\overline{h_i}$ as features. In other words, we estimate $p(y_i = 1|X_i) = \text{sigmoid}(\alpha + \beta_1 \hat{f}^{(\text{ext})}(X_i) + \beta_2 \overline{h_i})$.[4] The predicted probabilities from this logistic regression model constitute our final estimate of $p(y_i = 1|X_i)$; we use $\hat{f}(X_i)$ to refer to this estimate.

This two-stage procedure—first fitting a deep learning model to estimate the probability an image voted for Trump, and then estimating a logistic regression which combines the model output with human judgment—has three benefits:

(1) Inspecting the logistic regression coefficients allows us to *verify*, rather than *assuming*, that our model $\hat{f}^{(\text{ext})}(X_i)$ truly identifies ground-truth relevant signal that humans miss. If $\hat{f}^{(\text{ext})}(X_i)$ provided no additional signal beyond human judgment $\overline{h_i}$, the coefficient on $\hat{f}^{(\text{ext})}(X_i)$ in the fitted logistic regression model would be zero. Instead, the coefficient on $\hat{f}^{(\text{ext})}(X_i)$ is 22.2 (95% confidence interval, 20.5—24.0). The large and significant coefficient on $\hat{f}^{(\text{ext})}(X_i)$ indicates that the machine learning model is indeed detecting ground-truth relevant signal that humans miss and is thus useful for diagnosing human error.
(2) Our goal is to estimate the Bayes optimal model as accurately as possible, so we should use all features derivable

[4] We use a model with this simple parametric form because we find no evidence that more complex models—e.g., with interaction terms—improve our predictions.

	Train	Validation	Test
# of images	5,000	1,500	2,000
# of human responses	8,064,385	2,420,386	3,229,708
Accuracy of individual human	0.629	0.635	0.627
Accuracy of aggregate human	0.724	0.727	0.707
Responses per image (mean)	1,613	1,614	1,615
Responses per image (median)	1,611	1,613	1,614
Responses per image (std)	68.04	72.80	62.05
Responses per user ID (mean)	10.33	3.60	4.51
Responses per user ID (median)	8	3	3
Responses per user ID (std)	23.53	7.65	9.88
Fraction of images where $y_i = 1$	0.458	0.444	0.449
Fraction of images where $\overline{h_i} > 0.5$	0.563	0.557	0.567
Fraction of responses where $h_{ij} = 1$	0.537	0.535	0.538

Table 1: Descriptive statistics for the *New York Times* train, validation, and held-out test datasets. The differences across datasets for responses per user ID are expected, as the training set has approximately triple the number of images as do the other sets. All our main results are reported on the holdout test set (final column).

from the image, including human judgments based on the image. There is no guarantee that the deep learning model will capture all the signal on the image, given that it is trained on an external and finite dataset—although we do provide evidence below that our model is trained on a sufficiently large dataset for performance to level off, suggesting it approaches optimality. Substantiating this reasoning, the coefficient on $\overline{h_i}$ in our logistic regression is 2.0 (95% confidence interval, 1.5—2.4). The statistically significant coefficient on $\overline{h_i}$, though much smaller than that on $\hat{f}^{(\text{ext})}(X_i)$, indicates that humans in aggregate also pick up at least some signal that $\hat{f}^{(\text{ext})}(X_i)$ misses—likely because $\hat{f}^{(\text{ext})}(X_i)$ is trained on an finite external dataset which may not totally match the *New York Times* distribution. Combining both human judgment and the algorithmic prediction thus yields the best approximation of the Bayes optimal decision, which is our goal.

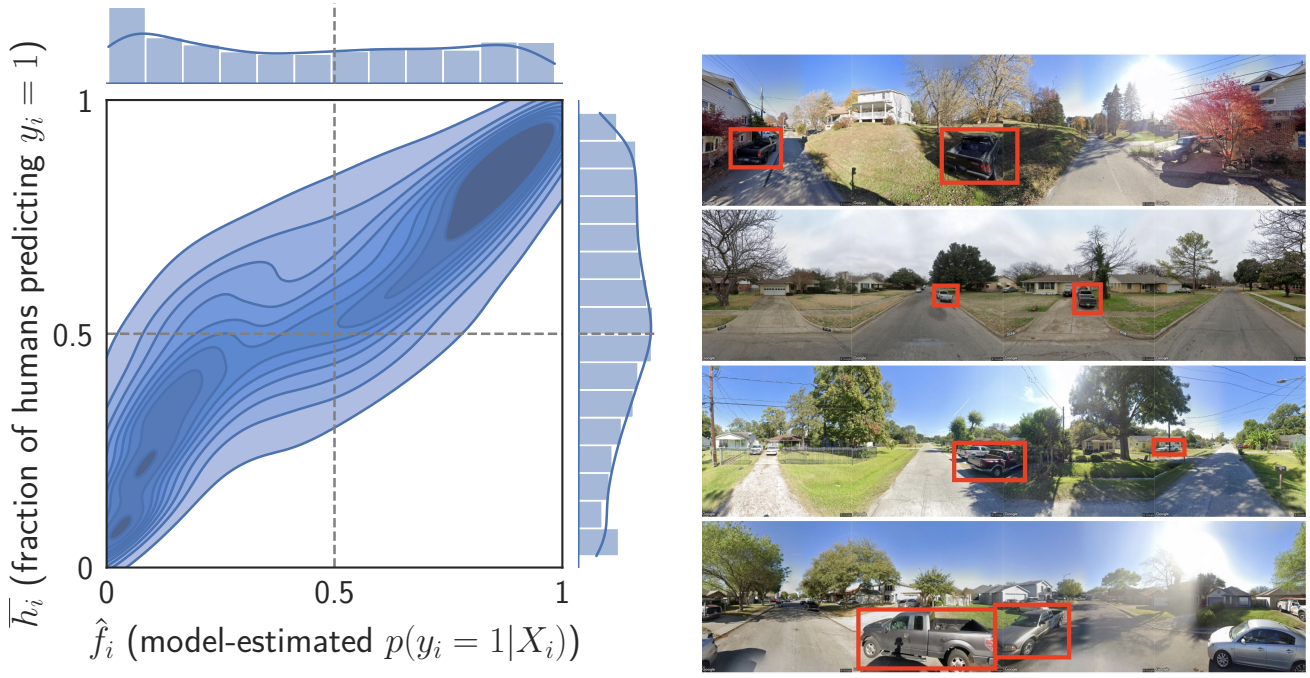
(3) Finally, as we discuss below, this two-stage estimation procedure will yield a useful lower bound on the magnitude of human bias even if the deep learning model we fit does not capture all the signal in the image.

Overall, a key strength of our two-stage approach is that it does not rely on being able to learn a machine learning model which perfectly estimates $p(y_i = 1|X_i)$ in order to provide useful insight into human error — often an impossible desideratum in small-data regimes. Rather, it merely requires that the machine learning model adds *some signal* beyond that captured in human judgment — something we can directly verify through logistic regression. Consistent with this, the accuracy of our final model $\hat{f}(X_i)$ in predicting ground truth y_i (75.1%) exceeds the accuracy of aggregate human judgment $\overline{h_i}$ (70.7%) or individual human judgment h_{ij} (62.7%).

Assessing the quality of the estimate of $p(y_i = 1|X_i)$. We verify two properties of our estimate of $p(y_i = 1|X_i)$. First, we show that it is calibrated (Figure A1) by comparing the model's predicted probabilities to the true fraction of positive examples for groups of observations binned by predicted probability, a standard check. Second, we verify that the machine learning model $\hat{f}^{(\text{ext})}(X_i)$ which forms a component of our final estimate of $p(y_i = 1|X_i)$ is trained on a sufficiently large dataset for model performance to level off, suggesting that we have enough training data that model performance is reasonably close to optimal: in Figure A2, we plot model performance with training sets of various sizes, showing that model performance levels off prior to our training set size. As we discuss below, one advantage of our two-stage approach is that it can still yield useful insights into human error even if $\hat{f}^{(\text{ext})}(X_i)$ is not completely optimal, as long as it adds signal beyond human judgment. Still, our approach will have more power to detect human error if the machine learning model it relies on is reasonably close to optimal performance, which is why we perform this check.

Assessing how $\overline{h_i}$ deviates from decisions implied by $\hat{f}(X_i)$. Next, we conduct a preliminary assessment of how human decision-making compares to what we would expect if humans were Bayes optimal. Figure 2a plots $\hat{f}(X_i)$, the model-estimated $p(y_i = 1|X_i)$, against $\overline{h_i}$. The relationship differs considerably from what we would expect if humans were Bayes optimal, in which case we would see a threshold function $\overline{h_i} = \mathbb{1}[\hat{f}(X_i) > 0.5]$. This implies that human decision-making is imperfect. Note that, unlike Figure 1, this figure suggests that humans are making *avoidable* errors—i.e., decisions which deviate from those of an estimated Bayes optimal decision-maker which has access to just the same images the humans do.

As a preliminary analysis of human error, we manually inspect individual images where $\overline{h_i}$ deviates particularly dramatically from decisions implied by $\hat{f}(X_i)$. Note that because $\hat{f}(X_i)$ is learned from both $\overline{h_i}$ and $\hat{f}^{(\text{ext})}(X_i)$, if $\hat{f}^{(\text{ext})}(X_i)$ added no additional signal above $\overline{h_i}$ for predicting ground truth, $\hat{f}(X_i)$ and $\overline{h_i}$ would be perfectly correlated, and there would be no dramatic deviations to examine at all; the presence of such deviations only occurs because $\hat{f}^{(\text{ext})}(X_i)$ does indeed add additional signal. We examine images where $\hat{f}(X_i)$ has high confidence that the neighborhood voted for Biden, but humans disagree—i.e., images in the top left corner of Figure 2a.



(a) (b)

Figure 2: (a) Plotting $\hat{f}(X_i)$, the model-estimated $p(y_i = 1|X_i)$, (x-axis) against aggregated human judgment $\overline{h_i}$ (y-axis) reveals that human judgment deviates substantially from Bayes optimality, which would produce a threshold function at $\hat{f}(X_i) = 0.5$. Instead, even for images where $fhat$ is only 25%, far more than 50% of respondents predicted Trump in some cases. Unlike Figure 1, this figure establishes that humans are making *avoidable* errors – ones not made by an approximately Bayes optimal decision-maker shown the same images the humans as human respondents. (b) 15 of the 21 images where humans most incorrectly skew towards Trump have pickup trucks (red bounding boxes), illuminating a source of human bias; we show 4 examples here; all 21 images also feature wide regions of open sky unobstructed by buildings. These images are identified by filtering for images with estimated $p(y_i = 1|X_i) < 0.2$ (Biden-leaning) and $\overline{h_i} > 0.6$ (Trump-leaning)—i.e., images in the top left region of (a). Underlying street view images ©Google.

(We set cutoffs at $\hat{f}(X_i) < 0.2$ and $\overline{h_i} > 0.6$, but our results are not sensitive to these thresholds.) 21 of the holdout test set images meet these criteria. Importantly, we find that of these images, only 19% in fact voted for Trump, indicating that $\hat{f}(X_i)$ is correct that these images are likely Biden neighborhoods and humans are in fact making avoidable errors. These images disproportionately have pickup trucks (trucks appear in 71% of these images as opposed to in 41% of the holdout test set as a whole[5]), indicating that humans believe trucks predict Trump more often than they really do (Figure 2b). This observation is indeed consistent with what *The New York Times* heard from some survey respondents, who in interviews said they believed that "pickup trucks were clear indications of a community's more conservative politics." All 21 images in this set also have wide regions of open sky unobstructed by buildings, another source of bias we discuss further in Section 4.3.

We identify no images where humans display the opposite bias—where the model is very confident that a neighborhood voted for Trump, $\hat{f}(X_i) > 0.8$, but humans disagree, $\overline{h_i} < 0.4$—indicating an interesting asymmetry in human judgment. Consistent with this, Table 1 shows that humans are slightly miscalibrated: they think neighborhoods vote for Trump more often than they really do[6]. Having established that simply examining the plot of how $\overline{h_i}$ deviates from decisions implied by $\hat{f}(X_i)$ can provide interesting insights into human error, we explore methods for more systematically decomposing this error below.

4.2 Bias-variance-noise decomposition of human error

Estimating $p(y_i = 1|X_i)$ also allows us to decompose human error into three sources—bias, variance, and noise—inspired by a classic decomposition for machine learning classifiers [27], and related to past work which seeks to assess both bias and variance in human decisions [40]. Prior to introducing our decomposition in the human

[5]In a random sample of 100 images from our holdout test set, manually inspected, we counted 41 images containing pickup trucks.

[6]Because we are interested in assessing all human biases, we do not calibrate human decisions prior to assessing them. However, calibrating $\overline{h_i}$ by choosing a threshold such that $\overline{h_i}$ classifies the correct number of images as Trump only increases its accuracy slightly (from 70.7% to 72.1%) likely because human decisions near the boundary are very noisy.

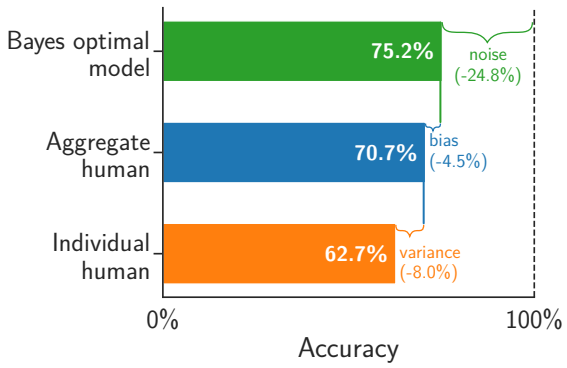


Figure 3: Sources of human error are decomposed into noise, bias, and variance terms by plotting the accuracy of our estimated Bayes optimal model, of aggregated human judgments, and of individual humans. (Pairwise differences between bars denote percentage point differences in accuracy.) As discussed in Section 4.2, *noise* is irreducible error that even the estimated Bayes optimal model cannot avoid; *bias* is additional error made by the aggregate human judgment for each image; and *variance* is the additional error due to disagreements between humans judging the same image.

decision-making setting, we briefly review the original decomposition in the machine learning setting [27]. Given a machine learning algorithm for learning a classifier, a training set of a fixed size, and a set of covariates, the goal of the decomposition is to assess why the machine learning algorithm performs imperfectly. [27] defines the *main prediction* as the aggregated prediction (e.g., majority vote in the case of zero-one loss) of classifiers fitted on different draws of the training set. Given this, the *bias* of the algorithm is the loss of the main prediction relative to the Bayes optimal prediction; the *variance* is the average loss of classifiers learned from individual training sets relative to main prediction; and the *noise* is the loss of the Bayes optimal classifier. In other words, the bias captures accuracy loss due to the inability of the model family to capture the true Bayes optimal model; the variance captures accuracy loss due to random variation across classifiers fitted on a finite train set; and the noise captures accuracy loss due to intrinsic unpredictability of the outcome from the features.

Our extension of this formalism to human decision-making is intuitive. We define the *main prediction* for humans as the binarized majority vote for each image, $\mathbb{1}[\overline{h_i} > 0.5]$, and conceptualize each human decision-maker as single fitted classifier from the "human model class". In Figure 3 we plot (1) the accuracy of the estimated Bayes optimal model $\hat{f}(X_i)$; (2) the human accuracy if every human agreed with the main prediction on each image; and (3) the accuracy of individual human predictions. The difference between (1) and perfect performance is the accuracy loss due to noise; the difference between (1) and (2) is the accuracy loss due to human bias; and the difference between (2) and (3) is the accuracy loss due to human variance. The accuracy losses due to variance (8.0 percentage points) and noise (24.8 percentage points) exceed those due to bias (4.5 percentage points). Thus, in this setting, our results establish that much of the error on the binary prediction task is unavoidable, and made even by our estimated Bayes optimal model: it is simply difficult to predict a neighborhood's election outcomes from a single Street View image. The fact that the accuracy of the aggregated human judgment $\overline{h_i}$ exceeds that of individual human judgments is consistent with previous work demonstrating wisdom of crowds [28]. If this were a real-world decision-making task, our results imply that having several humans judge each image would yield a considerable improvement over individual judgments, approaching the performance of a machine learning model. We note that if our model $\hat{f}(X_i)$ fails to capture the Bayes optimal model, we will overestimate the accuracy loss due to noise, and underestimate the loss due to human bias. Thus, our approach provides a useful *lower bound* on the magnitude of human bias even if our estimated model is not optimal.

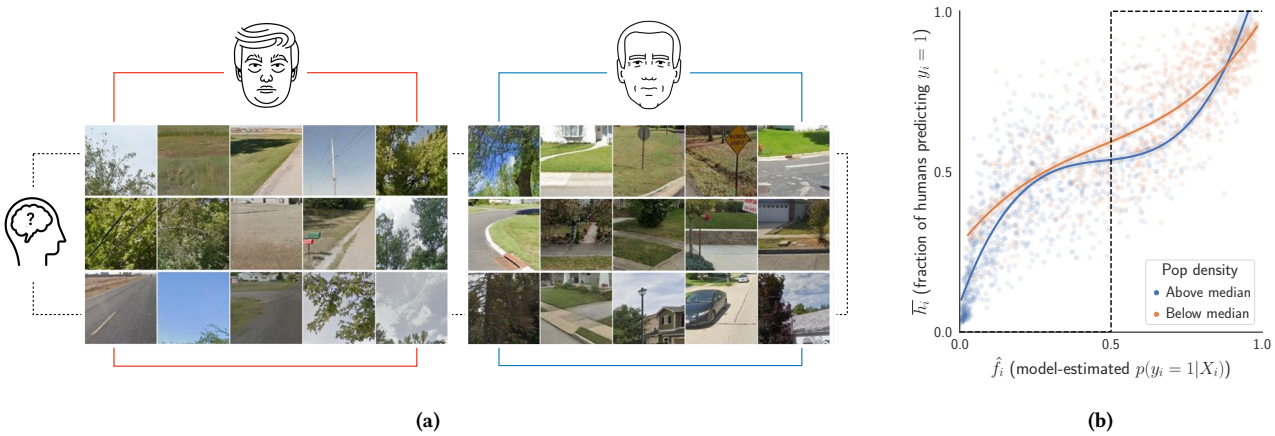
More broadly, we believe that our decomposition provides an actionable heuristic for assessing and improving decision-making processes in a wide variety of settings. For example, if doctors on a medical image classification task mainly lose accuracy due to bias, we may wish to consider retraining them or replacing them with an automated system; if they are accurate in aggregate but individually high-variance, we may need to solicit second opinions; and if the images themselves are noisy, we may need an alternate diagnostic modality.

4.3 What image features influence human judgment beyond the objective probability $p(y_i = 1|X_i)$?

Our manual inspection of images (Section 4.1) shows that deviations between $\overline{h_i}$ and $\hat{f}(X_i)$, our estimate of $p(y_i = 1|X_i)$, can provide insights about image features which lead humans astray, like pickup trucks (Figure 2b).

We now assess whether we can *systematically predict* from the image when $\overline{h_i}$ will deviate from what we would expect given our estimate of $p(y_i = 1|X_i)$: if there are image features that produce such systematic deviations, it suggests that humans are influenced by these features beyond what $p(y_i = 1|X_i)$ would justify. For example, consider our running example of pickup trucks, and suppose there is a pair of images with the same $p(y_i = 1|X_i)$—one with a pickup truck, and one without. If $\overline{h_i}$ for the image with the truck is greater than $\overline{h_i}$ for the image without the truck, this disparity is not justified by the objective probability of the image $p(y_i = 1|X_i)$: perhaps $\overline{h_i}$ for the truck image is too high, or $\overline{h_i}$ for the non-truck image is too low, but we can be sure that humans have a Trump-truck association beyond that justified by $p(y_i = 1|X_i)$. (Note that there may be some justified association between pickup trucks and probability that the neighborhood voted for Trump; however, this justified association would be captured in $p(y_i = 1|X_i)$).

To formalize this intuition, we train a second model to predict, from the image, how much $\overline{h_i}$ deviates from what we would expect given $p(y_i = 1|X_i)$. As before, we approximate $p(y_i = 1|X_i)$ with $\hat{f}(X_i)$. We do this in three steps:



(a) (b)

Figure 4: Density influences human decision-making beyond what is justified by $\hat{f}(X_i)$. (a) Image patches that caused the most shift in the residual model error towards Trump (left patches) or Biden (right patches). Patches of sky and other markers of low density, like road edges without curbs, on the left suggest that human respondents are more influenced towards Trump by features suggesting low population density than $\hat{f}(X_i)$ can explain. To further substantiate this observation, (b) shows how $\overline{h_i}$ varies as a function of $\hat{f}(X_i)$ for neighborhoods with higher than median population density (blue line) and lower than median population density (orange line). The orange line is higher than the blue line, indicating that, controlling for $\hat{f}(X_i)$, humans are more likely to think that less dense neighborhoods are Trump neighborhoods. The dotted line shows the Bayes optimal decision boundary. Underlying street view images ©Google.

(1) We first flexibly capture how $\overline{h_i}$ varies as a function of $\hat{f}(X_i)$ by fitting a cubic polynomial $\overline{h_i} = m(\hat{f}(X_i))$. $m(\hat{f}(X_i))$ corresponds to what the human average $\overline{h_i}$ tends to be for an image with estimated probability $\hat{f}(X_i)$.
(2) We then define the *residual* $r_i = \overline{h_i} - m(\hat{f}(X_i))$: that is, the portion of $\overline{h_i}$ that differs from how the aggregate human judgment tends to behave for images with the corresponding estimated probability $\hat{f}(X_i)$. If there are image features that predict this residual, this suggests that those image features lead to inconsistent human judgments about images with the same estimated probability $\hat{f}(X_i)$. We note that because we estimate $\hat{f}(X_i)$ using both $\overline{h_i}$ and $\hat{f}^{(\text{ext})}(X_i)$, if $\hat{f}^{(\text{ext})}(X_i)$ added no additional signal above $\overline{h_i}$ for predicting ground truth, $\hat{f}(X_i)$ and $\overline{h_i}$ would be perfectly correlated and the residual would be uniformly zero. Thus, in trying to predict a non-zero residual, we are attempting to predict signal which truly arises from the fact that the machine learning model $\hat{f}^{(\text{ext})}(X_i)$ identifies ground-truth relevant signal which humans miss.
(3) To search for the image features which predict the residual, we train a neural network to predict the residual from the image: $\hat{r}_i = g(X_i)$ (see Appendix A.2 for model training details).

We find that the neural network is able to achieve statistically significant signal for predicting the residual from the image (Spearman correlation between true and predicted residual, 0.534; $p < 0.001$). This statistically significant correlation shows that r_i is systematically predictable from the image, indicating that there are image features which cause humans to deviate systematically from consistent responses to the estimated objective probability $\hat{f}(X_i)$.

Interpreting the residual model. To identify the specific image features which cause this systematic deviation in the residual, we use occlusion mapping [67] to interpret the fitted residual model. Specifically, we identify the image patches which most change the residual model's predictions when they are masked out (Appendix A.3). The results of this analysis are shown in Figure 4a, illustrating that patches of sky and other features indicating low population density push the residual model in the Trump direction: in other words, human respondents are more swayed towards Trump by visual indicators of low population density than the estimated objective probability can explain.[7] As further evidence, the correlation between $\hat{r}_i$ and log population density is also negative (Spearman r of -0.234, $p < 0.001$): Figure 4b illustrates that, controlling for the estimated objective probability, humans think that denser neighborhoods are more likely to be Biden neighborhoods. Overall, this analysis shows that humans are swayed by population density beyond what our estimate of $p(y_i = 1|X_i)$ can explain: given two images with identical estimated $p(y_i = 1|X_i)$, humans will be more likely to think the denser neighborhood voted for Biden. Interviews conducted by *The New York Times* confirm that some readers did indeed use density to guide their decision-making [7].

5 ALTERNATIVE APPROACHES

In developing the method described in the prior section, we also considered two alternative methods for diagnosing human error in

[7] The open sky patches in Figure 4a are consistent with the open skies seen in individual images in Figure 2b, indicating that the manual inspection of individual images is yielding conclusions consistent with the quantitative residual analysis.

image analysis, and describe them here—concluding that though they are intuitive they each suffer from drawbacks which render our primary approach preferable.

5.1 Alternative approach 1: Train two models to predict ground truth and human judgment

A straightforward algorithmic approach to diagnosing human error is to train one model to predict human judgment and a second model to predict ground truth, arguing that discrepancies between the two models indicate human error. To investigate this approach, we train one model to predict the aggregate human judgment $\overline{h_i}$, achieving an RMSE of 0.09 and a Spearman r of 0.93; we train a second model to predict the ground truth continuous vote difference $y_i^{(\text{cont})}$, achieving an RMSE of 0.30 and a Spearman r of 0.67 (Appendix A.2).[8] We study how the two models differ by comparing the image patches which most change model predictions when they are masked out using the same occlusion mapping technique described in Section 4.3. This method is direct and intuitive, and we show the results from it in Figure A4—revealing, for example, that the ground truth model associates road patches more strongly with Trump than does the human judgement model. In particular, two-lane highways divided by double-yellow lines are the most predictive of Trump neighborhoods, possibly because these highways signal the area is more rural.

However, this approach has several downsides. First, systematically comparing the image regions which influence two different deep learning models is difficult; identifying and interpreting salient features for even a *single* model is a subtle and active area of research [3]. For example, if both models appear to be influenced by trucks, but the model predictions change by different amounts when trucks are occluded, it is unclear whether humans are misweighting trucks or the scales of the two model targets are simply incomparable. Second, this approach requires a ground truth model which outperforms human judgment; if humans perform better than the ground truth model, it is hard to argue that deviations from it are human mistakes. (Our main approach has a similar requirement: if humans outperform our estimated model of $p(y_i = 1|X_i)$, it is hard to argue that we are correctly identifying human error. However, we meet this requirement by design, by including the average human judgement as a feature in our model of $p(y_i = 1|X_i)$, and we verify that we outperform human judgment.) In this setting, we have enough data to train models that outperform human judgment, but in small data settings this may be difficult to do.

5.2 Alternative approach 2: predict the difference between ground truth and human judgment

A second option is to train a model to directly predict the difference between ground truth and human judgment, $d_i = \overline{h_i} - y_i$, and then use occlusion mapping to identify image regions which contribute to a large difference. This method has several advantages over the method described in Section 5.1: it is simpler, and it doesn't require a model which outperforms human judgment—just one which can predict the difference between ground and human judgment. However, it suffers from a conceptual flaw: this difference will be predictable even if humans are Bayes optimal—and so this method would incorrectly identify image features as causing errors even if humans are using them optimally. For example, suppose the only informative feature is whether the image has a car in it, and that 70% of images with cars vote Trump while 20% of images without cars vote Trump. Then, Bayes optimal humans will always classify images with cars as Trump (so $d = 0.3$ on average on car images) and images without cars as Biden (so $d = -0.2$ on average on non-car images). Our model will learn that cars predict the difference, implying that humans are over-weighting cars—even though in fact humans are Bayes optimal, with the error stemming from the fact that the images are not sufficiently informative. We view this conceptual flaw as sufficiently serious that we do not present results from this approach. We note that the method we favor in Section 4.3, which fits a model to estimate $r_i = \overline{h_i} - m(\hat{f}(X_i))$ as opposed to $d_i = \overline{h_i} - y_i$, overcomes this limitation: if humans are Bayes optimal, r_i will be uniformly 0 and we will not be able to identify image features which correlate with it. Observing this conceptual flaw was a primary motivation for our method and in particular in developing a model to estimate $p(y_i|X_i)$.

We note that our favored method avoids the major weaknesses of both alternate approaches described above. To avoid having to compare occlusion maps between two different models, a weakness of the first alternate method described in Section 5.1, our favored method examines the occlusion map from only a single model trained to predict the residual $\overline{h_i} - m(\hat{f}(X_i))$. To avoid the conceptual mistake of predicting the difference $\overline{h_i} - y_i$, a weakness in the second alternate method described in Section 5.2, our favored method predicts the difference $\overline{h_i} - m(\hat{f}(X_i))$ rather than $\overline{h_i} - y_i$.

6 DISCUSSION

In this work, we use a unique dataset of over 16 million human judgments with ground truth to propose a method for diagnosing human error in image analysis, a uniquely challenging setting for diagnosing human error. We show that by estimating $p(y_i = 1|X_i)$, we can decompose human error into bias, variance, and noise terms, and also identify specific image features which influence human judgment beyond the objective probability of the image $p(y_i = 1|X_i)$. We show that even if the machine learning model which forms a component of our estimate of $p(y_i = 1|X_i)$ is not perfectly optimal — which will frequently be true for models trained on complex inputs like images especially in small-data regimes — our approach can still provide useful insights into human error as long as the model adds signal beyond human judgment, a property we verify. We consider two alternate methods for diagnosing human error and assess their flaws. To facilitate reproduction and extension of our results, code to implement our method and reproduce our results is publicly available at https://github.com/zamfi/diagnosing-human-error-in-image-analysis.

Limitations. There are several caveats in interpreting our results. First, we interpret our convolutional neural network models using

[8] It is intuitive that prediction performance for $\overline{h_i}$ is better, because it is easier to predict: $\overline{h_i}$ should be almost entirely determined by the image, whereas it is unlikely that $y_i^{(\text{cont})}$ is.

standard and widely used interpretability techniques, but these are known to sometimes yield misleading conclusions [3, 39]. Second, our data comes from self-selecting respondents to the *New York Times* quiz, and as such the specific patterns we observe may not generalize to other populations. Although the methods we develop apply much more generally, any observed bias is only as representative as "human bias" as our sample of humans is representative of all humans. Third, our decomposition of human error into bias, variance and noise terms relies on our estimate of $p(y_i = 1|X_i)$ being approximately optimal, and we otherwise provide a *lower bound* on the magnitude of human bias. While we provide suggestive evidence that our estimate of $p(y_i = 1|X_i)$ is reasonably close to optimal, we cannot verify this conclusively. Fourth, we cannot conclusively say that human error is due solely to the contents of the images humans are asked to evaluate; systematic errors can also be caused by poor task instructions or confusing user interface design, for example, and our method cannot isolate these effects specifically. Explicit variation in instructions or user interface design, if tracked, could serve as an additional variable to consider alongside or in conjunction with our method.

Future directions

There are many potential directions for future work. Methodologically, there are several potential extensions to our current method. First, human decision-makers are heterogeneous: for example, previous work has developed methods for clustering humans by the errors they make [44]. It would be interesting to extend our method to model heterogeneity in human decision-makers. Second, we focus on the setting where ground truth labels are available for all observations, but in many real-world settings, "selective labels" mean that ground truth is censored by human judgments [40, 54]: for example, we only observe a test result if a doctor decides to order a test. Extending our method to accommodate selective labels settings represents another avenue for future work.

Another direction for future work is applying our approach to other datasets. First, there are many other image datasets where our method could be applied: the ideal use case for our approach is an image dataset with human judgments h_{ij} and an objective ground truth label y_i which is defined independently of human judgment. There is increasing recognition in the machine learning community that such objective ground truth labels (e.g., mortality in a medical setting) are invaluable to avoid merely laundering human biases into "ground truth" [16], and datasets are correspondingly becoming more widely available: for example, the Nightingale Open Science initiative[9] is an effort to collect and make publicly available such datasets. Second, while we develop our method for image data, a more challenging setting than tabular data, many of our insights could equally be applied to tabular datasets—for example, the bias-variance-noise decomposition for human error we propose. Finally, our method is in principle applicable not just to human judgments, but to black-box algorithmic decision-making systems as well. We believe the method we propose is broadly applicable to diagnose and dissect both human and algorithmic error in a wide variety of settings.

[9]https://docs.nightingalescience.org/

ACKNOWLEDGMENTS

The authors would like to thank *The New York Times* for providing the initial dataset enabling this work and Michael Elabd for his contributions to an earlier version of this project. They thank Shengwu Li, the Cornell AI Policy and Practice group, and the anonymous reviewers for their thoughtful comments. Emma Pierson was supported by a Google Research Scholar Award and J.D. Zamfirescu-Pereira was partially supported by the United States Air Force and DARPA under contracts FA8750-20-C-0156, FA8750-20-C-0074, and FA8750-20-C0155 (SDCPS Program). These funders had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 252–260.

[2] Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M Small, and H Andrew Schwartz. 2019. Using search queries to understand health information needs in africa. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 3–14.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292* (2018).

[4] Mayank Agrawal, Joshua C. Peterson, and Thomas L. Griffiths. 2020. Scaling up psychology via Scientific Regret Minimization. *Proceedings of the National Academy of Sciences* 117, 16 (2020), 8825–8835. https://doi.org/10.1073/pnas.1915841117

[5] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[6] Ashton Anderson, Jon Kleinberg, and Sendhil Mullainathan. 2017. Assessing human error against a benchmark of perfection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 4 (2017), 1–25.

[7] Emily Badger, Josh Katz, and Kevin Quealy. 2021. What We Learned From 15 Million Guesses About a Neighborhood's Politics. *The New York Times* (2021).

[8] Emily Badger, Josh Katz, Kevin Quealy, and Rumsey Taylor. 2021. Do you think you can tell how a neighborhood voted just by looking around? *The New York Times* (2021).

[9] Gary S Becker. 2010. *The economics of discrimination*. University of Chicago press.

[10] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine* 15, 11 (2018), e1002699.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.

[12] Michael A Bruno, Eric A Walker, and Hani H Abujudeh. 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics* 35, 6 (2015), 1668–1676.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[14] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[15] Colin F. Camerer. 2019. *24. Artificial Intelligence and Behavioral Economics*. University of Chicago Press, 587–610. https://doi.org/doi:10.7208/9780226613475-026

[16] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4 (2020).

[17] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[18] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[19] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 598–609.

[20] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[21] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491* (2014).

[22] Robyn M. Dawes. 1971. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist* 26, 2 (1971), 180–188.

[23] Robyn M. Dawes, David Faust, and Paul E. Meehl. 1989. Clinical Versus Actuarial Judgment. *Science* 243, 4899 (1989), 1668–1674. https://doi.org/10.1126/science.2648573

[24] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. *A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[25] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 2970–3005. https://doi.org/10.18653/v1/N19-1304

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[27] Pedro Domingos. 2000. A unified bias-variance decomposition for zero-one and squared loss. *AAAI/IAAI* 2000 (2000), 564–569.

[28] Francis Galton. 1907. Vox populi (the wisdom of crowds). *Nature* 75, 7 (1907), 450–451.

[29] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[30] Sharad Goel, Justin M Rao, and Ravi Shroff. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics* 10, 1 (2016), 365–394.

[31] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. https://doi.org/10.1145/3359152

[32] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022). https://doi.org/10.1073/pnas.2110013119 arXiv:https://www.pnas.org/content/119/1/e2110013119.full.pdf

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[34] Sophie Hilgard, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, and David Parkes. 2021. Learning Representations by Humans, for Humans. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 4227–4238. https://proceedings.mlr.press/v139/hilgard21a.html

[35] Todd C Hollon, Balaji Pandian, Arjun R Adapa, Esteban Urias, Akshay V Save, Siri Sahib S Khalsa, Daniel G Eichberg, Randy S D'Amico, Zia U Farooq, Spencer Lewis, et al. 2020. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine* 26, 1 (2020), 52–58.

[36] Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. 2018. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651* (2018).

[37] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[38] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: a flaw in human judgment.* Little, Brown.

[39] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 267–280.

[40] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[41] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018), 113–174.

[42] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2020. Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30096–30100.

[43] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[44] Himabindu Lakkaraju and Jure Leskovec. 2016. Confusions over Time: An Interpretable Bayesian Model to Characterize Trends in Decision Making. In *NIPS*. 3261–3269.

[45] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[46] Emma Lurie and Eni Mustafaraj. 2019. Opening Up the Black Box: Auditing Google's Top Stories Algorithm. In *The Thirty-Second International Flairs Conference.*

[47] David Madras, Toniann Pitassi, and Richard Zemel. 2017. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664* (2017).

[48] Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2020. Aligning Superhuman AI with Human Behavior: Chess as a Model System. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1677–1687.

[49] Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2020. Learning personalized models of human behavior in chess. *arXiv preprint arXiv:2008.10086* (2020).

[50] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[51] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[52] Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*. PMLR, 7076–7087.

[53] Sendhil Mullainathan. 2019. Biased algorithms are easier to fix than biased people. *The New York Times* (2019).

[54] Sendhil Mullainathan and Ziad Obermeyer. 2021. Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *The Quarterly Journal of Economics* (12 2021). https://doi.org/10.1093/qje/qjab046 qjab046.

[55] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. 2018. Fast threshold tests for detecting discrimination. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 96–105.

[56] Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. 2021. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine* 27, 1 (2021), 136–140.

[57] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature human behaviour* 4, 7 (2020), 736–745.

[58] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*. PMLR, 5281–5290.

[59] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014), 4349–4357.

[60] Megan T Stevenson and Jennifer L Doleac. 2021. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440* (2021).

[61] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[62] F Tan, D Caicedo, A Pandharipande, and M Zuniga. 2018. Sensor-driven, human-in-the-loop lighting control. *Lighting Research & Technology* 50, 5 (2018), 660–680. https://doi.org/10.1177/1477153517693887

[63] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[64] André Calero Valdez and Martina Ziefle. 2018. Human factors in the age of algorithms. understanding the human-in-the-loop using agent-based modeling. In *International Conference on Social Computing and Social Media*. Springer, 357–371.

[65] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

[66] Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al. 2021. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digital Medicine* 4, 1 (2021), 1–13.
[67] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.
[68] Quanshi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *arXiv preprint arXiv:1802.00614* (2018).

A APPENDIX

A.1 Data augmentation

To better train our machine learning models, we augmented the *New York Times* dataset with 52,025 additional images, alongside their ground truth vote shares.

The official *New York Times* dataset was constructed by randomly selecting 10,000 voter addresses for which precinct-level results were available, ensuring the sample was representative of the 2020 vote in both vote margin and population density [8]. We replicate this methodology to acquire 52,025 additional images as follows, with the goal of producing an expanded dataset as similar as possible to the original *New York Times* dataset (and indeed, we confirm as a robustness check that models trained on the original dataset yield similar performance on the expanded dataset).

(1) We start with a list of all electoral precincts, their geographic shapefiles, and their 2020 election results, compiled by *The New York Times*: https://github.com/TheUpshot/presidential-precinct-map-2020. We draw a sample of precincts which matches *The New York Times* sample on vote margin and population density because these were the variables used to rebalance *The New York Times* sample.
(2) Using a proprietary voter file made available to us by an election analytics firm, for each precinct we sample up to 10 voters whose home address lies in the precinct.
(3) We use the Google Street View API to retrieve Street View images for each home location.
(4) Finally, to compensate for bias introduced by querying the Street View API (since not all home locations have Street View images), we again resample our dataset so it matches the original *New York Times* data on vote margin and population density. (We confirm that the two datasets are also similar on other census demographics features like median age, household income, race/ethnicity, education, insurance levels, and home-ownership.)

A.2 Training image models

We train deep learning models to predict three targets from Street View images: the aggregated human judgment $\overline{h_i}$, the continuous ground truth vote margin $y_i^{(\text{cont})}$, and the binary ground truth election outcome y_i.

$\overline{h_i}$ model. To train a model to predict $\overline{h_i}$ from a neighborhood image, we begin with a ResNet model as the base model [33]. Because each neighborhood X_i is represented by four images (corresponding to viewing angles of 0°, 90°, 180°, and 270°), our model architecture passes the images individually through the pre-trained ResNet and uses the final feature layer as the representation of each image. We then concatenate these representations to obtain a complete representation of all four images, which is then used as input into a series of fully-connected and ReLU layers to predict $\overline{h_i}$. We initialize the model with weights pre-trained on ImageNet [26] and fine-tune the model on our dataset. We perform hyperparameter search over the ResNet architecture (ResNet-34, ResNet-50, ResNet-101, or ResNet-152), the proportion of layers in the base ResNet model to unfreeze, optimizer parameters such as learning rate and decay, the resolution of the input images, and whether to randomly flip and crop the inputs. We train the $\overline{h_i}$ models on the *New York Times* data with a mean squared error loss function and select the model with the lowest loss on the *New York Times* validation dataset. Our model achieves 88.0% accuracy on the holdout test set when $\overline{h_i}$ and the model predictions are binarized at 0.5, a Spearman r of 0.93, and an RMSE of 0.09.

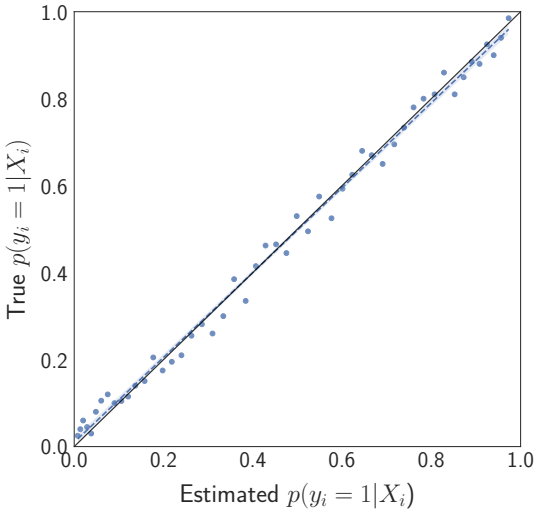


Figure A1: The model-estimated calibrated probabilities $\hat{f}(X_i) = p(y_i = 1|X_i)$ (x-axis) line up well with the true probabilities, demonstrating that the model is calibrated. Observations are divided into 50 bins, sorting by $\hat{f}(X_i)$; each point compares the mean values of $\hat{f}(X_i)$ and y_i in one bin.

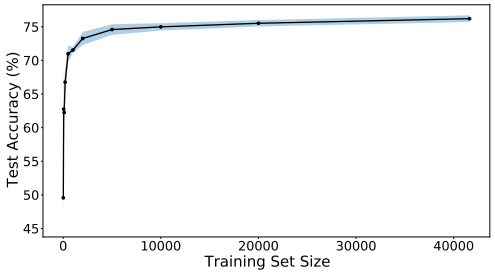


Figure A2: Performance of $\hat{f}^{(\text{ext})}(X_i)$ on training sets of different sizes. To reduce noise for small train sets, accuracy for each train set size is averaged across five randomly drawn train sets. Errorbars show the standard deviation across the five iterations. Model performance levels off as we approach the full train set size, suggesting that the train set size is large enough for model performance to approach optimality.

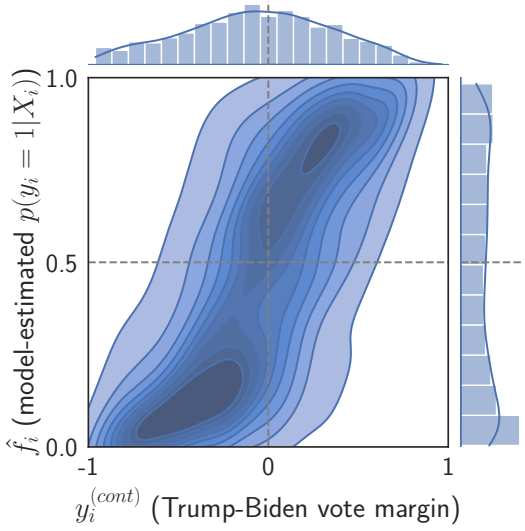


Figure A3: The model-estimated calibrated probabilities $\hat{f}(X_i) = p(y_i = 1|X_i)$ (y-axis) are positively correlated, as expected, with the actual vote share $y_i^{(cont)}$ (x-axis): the precincts where the vote was close are also those where the model expresses the greatest uncertainty.

$y_i^{(cont)}$ *model.* Due to the similarity between the tasks of predicting $\overline{h_i}$ and $y_i^{(cont)}$—both are predicting a continuous output from the neighborhood image as input—the model architecture and hyperparameter search remain the same. However, we train the $y_i^{(cont)}$ models using the larger external training and validation sets before selecting the best model using the *New York Times* validation set. (We cannot do this for the $\overline{h_i}$ model because we do not have data on $\overline{h_i}$ for the external dataset.) The highest-performing model that estimates $y_i^{(cont)}$ achieves 74.6% test accuracy when binarized at a threshold of 0, in other words predicting whether a neighborhood voted for Biden or Trump. In addition, the Spearman correlation between the predicted and true $y_i^{(cont)}$ on the test set is 0.69, $p < 0.001$, and the RMSE is 0.29.

$p(y_i = 1|X_i)$ *model.* The model, training, and dataset setup to estimate $p(y_i = 1|X_i)$ are largely identical to that of the $y_i^{(cont)}$ model. However, because y_i is binary, we employ an additional sigmoid layer to convert the unbounded continuous output to a probability between 0 and 1. We also treat the task as a binary classification task to predict whether input neighborhoods voted for Biden (0) or Trump (1), rather than a continuous prediction task, and therefore use a negative log-likelihood loss function rather than MSE loss. Our classifier achieves a test accuracy of 74.0% and

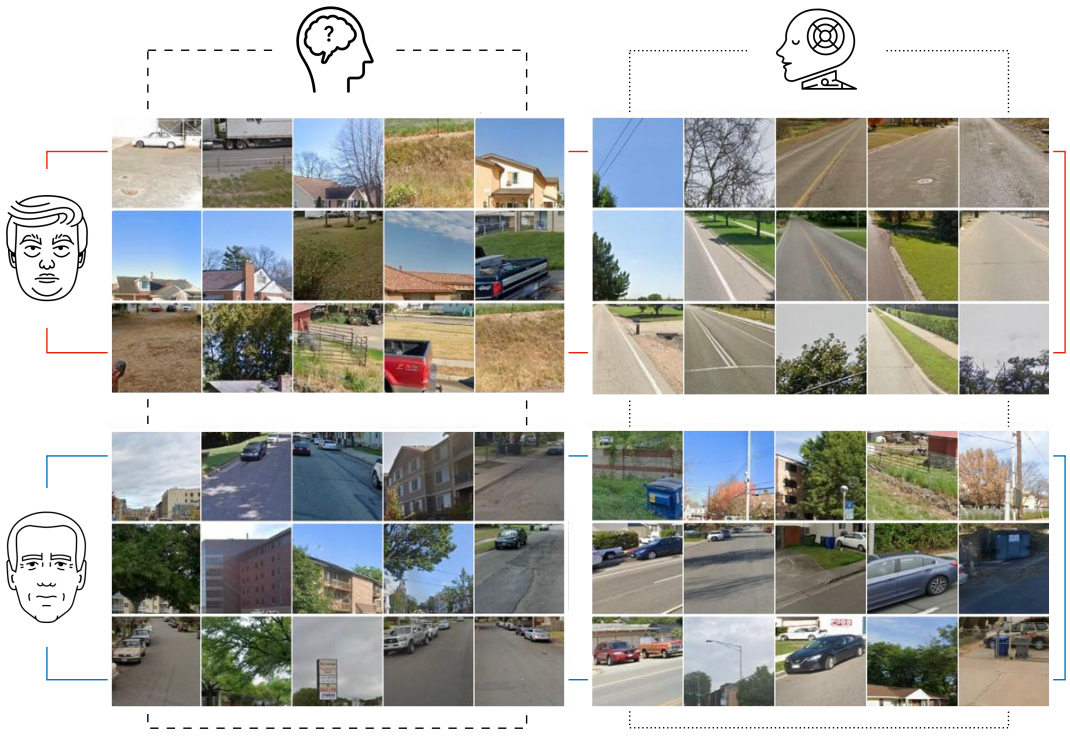


Figure A4: Image patches that most shift the model prediction towards Trump (top row) or Biden (bottom row), for the model which predicts the aggregate human judgment $\overline{h_i}$ (left column) and the model which predicts ground truth $y_i^{(cont)}$ (right column). There is a clear visual difference between the left and right columns—for example, roads figure more prominently in the top right than the top left—indicating a difference in the features that most influence the ground truth model and the human judgment model. Underlying street view images © Google.

an AUC of 0.82. A density plot comparing our modeled $p(y_i = 1|X_i)$ with the ground truth $y_i^{(\text{cont})}$ appears in Figure A3.

r_i model. Similarly, the model, training, and dataset setup to produce $\hat{r}_i$ are largely identical to that of the $\overline{h_i}$ model, except that the target for training is the residual $r_i = \overline{h_i} - m(\hat{f}(X_i))$ and we use the *New York Times* training and validation datasets for model training and selection. We cannot use the external dataset to train the residual model because we do not have data on $\overline{h_i}$.

A.3 Identifying image regions most influencing prediction

To identify image regions which influence a model's predictions (as in Figure A4) we mask out regions of the image and measure the change in model predictions, following previous work [67]. Specifically, we divide each of the 0°, 90°, 180°, and 270° Google Street View angles into a 4x4 grid, yielding a total of 64 square regions for each neighborhood image; for each square region, we measure how much the model prediction changes when we replace the square with a 60% gray square. This yields a value for each square region which captures the impact of the region on the model's prediction. (We verify that the the results we report, e.g., finding "open skies", are robust to using a 2x2 grid size instead.)

In Figure A4, we show the regions which produce the largest changes in model outputs for the models predicting $\overline{h_i}$ and $y_i^{(\text{cont})}$, respectively. Comparing these regions corresponds to the method described in Section 5.1.