A Appendix A: Conceptual Framework

A.1 A Simple Framework for Discovery

In this section, we develop a simple framework to clarify the goals of hypothesis generation, how it differs from testing, and how algorithms can help. The goal is to organize thinking around these largely unmodeled questions rather than to prove intricate, formal results. The focus of our model—and our work—is the act of generating hypotheses given some data about the world, as opposed to, say, deriving hypotheses from formal models.

A.2 Setup

We assume the overall goal is to understand how some outcome (y) depends on some input (x). Simply finding any function from x to y is not enough. Instead, for us, understanding requires that we uncover "interpretable" functions that relate the two. The goal of hypothesis generation is to uncover candidate interpretable functions. Testing—which we do not model—happens next: It verifies whether a given candidate function in fact explains the relationship between y and x. The hypotheses are to be generated from some data, which we assume consist of (y, x) pairs that are from a distribution $\mathcal{D}$. We will also assume that $x \in \mathbb{R}^k$ is a k-dimensional real vector, $y \in \{0, 1\}$ is some binary outcome, and that there is some function $f^* : \mathbb{R}^k \to \{0, 1\}$ such that $y = f^*(x)$ for all (x, y) sampled from $\mathcal{D}$.[72]

Though our framework is general, to make it concrete, we will describe it using the application we study in the rest of the paper: x is a mugshot (a picture of an arrested individual's face) and y is the judge's decision to detain a defendant. Images can be represented as a vector of pixel values, where each element of the vector indicates the intensity of a given color (red, green or blue) at a particular point in the image. A $1,024$ by $1,024$ pixel image could be expressed as a vector of length $3,145,728$ $(1,024 \times 1,024 \times 3)$. So each mugshot x is an element of $\mathbb{R}^k$, the space of all possible pixel combinations. A variety of other inputs can be represented similarly. Suppose a diary entry (x) is related to whether a person is depressed or not (y). The entry can be "vectorized" by replacing each word with an ID. The entry can be represented with a series of concatenated word IDs followed by a series of dummy vectors to make sure all vectorized entries have similar length.

In this framework, we are not interested in the estimates produced by f^*, but rather in uncovering "interpretable" functions that explain how x relates to y. For example, skin color and attractiveness (as judged by a certain group) are interpretable functions that can be calculated for a given image x. In the diary example, interpretable functions might include a measure of whether the text is "sad", or if it "discusses self more than others," or "discusses emotions"; all of these are interpretable functions of the input vector which have some meaning to people. But the converse is not true: Most mathematical functions of images are not interpretable, and have no meaning to us. With this in mind, let $\mathcal{H}$ be the set of all possible hypotheses—the set of every function from input data x to the chosen outcome y. We assume there is also some set $\mathcal{I} \subseteq \mathcal{H}$ of *interpretable* functions, which are the functions that admit some human-understandable description.[73] We also assume for a given

[72] Relaxing the simplifying assumptions that y is binary and that there is no noise in the relationship between x and y does not substantively change the setup or the substance of our results.

[73] Note the set $\mathcal{I}$ is not static: Fundamental innovations such as the discovery of calculus or probability can

data distribution $\mathcal{D}$ with associated function f^*, there is a set of comprehensible functions,

$$\Phi^* = \{\phi_1^*, ...\phi_j^*\} \subset \mathcal{I},$$

that together explain the human interpretable component of the ground-truth. That is, f^* can be written as

$$f^*(x) = \underbrace{\phi_1^*(x) + \phi_2^*(x) + \cdots + \phi_j^*(x)}_{\text{Interpretable}} + \Delta^*(x). \tag{1}$$

Reality may be interpretable, but need not be, hence $\Delta^*(x)$. The goal of discovery is to find a candidate set of comprehensible functions $\Phi \subset \mathcal{I}$, such that

$$f_\Phi(x) = \sum_{\varphi \in \Phi} \varphi(x)$$

is as close to f^* as possible, as measured by some error function such as likelihood (or mean-squared error for continuous outcomes). Finding such Φ is desirable partly because interpretable insights are portable. Learning that skin color predicts detention has broader implications: We may now want to ask whether skin color affects police use of force or whether these effects differ by time of day. By virtue of being interpretable, the functions in $\mathcal{I}$ let us use a wider set of knowledge (police may share racial biases, or skin color is not as easily detected at night). We seek interpretable descriptions because they let us generalize to novel situations, in addition to being easier to communicate to key stakeholders and lend themselves to interpretable solutions. The interpretable set here models the set of ideas for which we have some broader understanding, which in turn lets us perform these acts of generalization.

It is worth pointing out what makes this task hard (or easy). Suppose we simply built a model $m(x)$ that predicts y. For that model itself to yield an interpretable hypothesis, its parameters must be interpretable. That can happen in some simple cases. For example, if we had a data set where each dimension of x was interpretable (such as individual structured variables in a tabular dataset) and we used a predictor such as OLS (or LASSO), then we could just read the hypotheses from the non-zero coefficients: which variables are significant? Even in that case, interpretation is challenging because machine learning tools, built to generate accurate predictions, yield coefficients that can be quite unstable and change with small perturbations in the data (Mullainathan and Spiess, 2017).[74] And often interpretation is much less straightforward than that. If x is an image, text or time series, the estimated models (such as convolutional networks) are defined on granular inputs and have no particular meaning: if we knew the algorithm weighted a particular pixel, what have we actually learned? In these cases, the estimated model m is itself not interpretable, nor is it readily apparent how we might decompose m into interpretable sub-components in the same way that we have decomposed f^* in (1). Our focus is on these contexts where algorithms, as "black-box" models, are not readily interpreted.

---

expand the set of functions that "make sense" to us. But for present purposes we take $\mathcal{I}$ as given, and do not focus on ground-breaking discoveries that truly expand our basic capacity to represent the world.

[74] The intuition here is quite straightforward: If two predictor variables are highly correlated, the weight that the algorithm puts on one versus the other can change from one draw of the data to the next depending on the idiosyncratic noise in the training dataset, but since the variables are highly correlated the predicted outcome values themselves (hence predictive accuracy) can be quite stable.

A.3 The Discovery Problem

To understand the challenge of making such discoveries, we first define the process of extracting candidate hypotheses from a given dataset. Expressed in the language of our framework, a discovery process $\mathcal{P}$ takes a data set D as input, and returns a function h as output. The output function h is our candidate hypothesis. The process $\mathcal{P}$ may be random, meaning that hypotheses drawn from $\mathcal{P}(D)$ may be different each time. Having drawn a candidate hypothesis h we can then evaluate h as a candidate element of Φ, but this is a separate problem to discovery. We call $\mathcal{P}$ a *discovery procedure*, and we call h the *hypothesis*.

Under our definition, each of the following is considered a discovery procedure: a researcher fitting a linear regression; fitting a black-box machine learning model to a dataset (the model is the hypothesis); or a researcher arriving at a hypothesis by manually inspecting a dataset and having a creative inspiration. So we need to define what distinguishes a "good" discovery process from a poor one. We introduce three properties of a discovery process.

Definition. *Suppose that we have a distribution $\mathcal{D}$ over data sets, and a discovery procedure $\mathcal{P}$ for data sets drawn from $\mathcal{D}$. Let D be a single data set drawn from $\mathcal{D}$, and let $h = \mathcal{P}(D)$ be a hypothesis drawn from the discovery procedure. We define three criteria for measuring the the quality of $\mathcal{P}$.*

Comprehensibility*: The generated hypotheses should be ones that people can understand. Comprehensibility is defined as*

$$\xi(\mathcal{P}) = \mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}).$$

Plausibility*: Under the data-generating distribution, the generated hypothesis should at least be predictive of y. Plausibility is defined as*

$$\pi(\mathcal{P}) = \mathbb{P}_{\mathcal{D}}\Big(\operatorname{cor}(h(x), y) > 0 \Big).$$

Replicability*: Repeating the discovery procedure should lead to the same hypothesis. Let D' be another data set drawn from $\mathcal{D}$, and let $h' = \mathcal{P}(D')$ be another hypothesis drawn from the discovery procedure on the new data set. Replicability is defined as*

$$\rho(\mathcal{P}) = \mathbb{P}_{\mathcal{D}}(h = h').$$

For all three criteria, probabilities and expectations are taken over new draws from the original data-generating process $\mathcal{D}$.[75]

We note four points about these criteria. First, why do we include replicability? It is hard to systematically assess a procedure that is highly idiosyncratic. We would not be able to assess the quality of a procedure, or whether a produced hypothesis was due to the procedure or simply a random draw (hence the problem with human creativity). Second, does plausibility necessarily and mathematically imply replicability? If a procedure only finds plausible candidates, is it not by construction replicable? The extent to which this is true depends on the size of the plausible set—if there are few plausible hypotheses, high

[75] As a result, the performance of a given procedure can depend on $\mathcal{D}$, which is intuitive: some procedures may do better for some data-generating processes than others. Additionally, We could have included another criterion here: *novelty*. Because some of the hypotheses $\phi_i^*(x)$ may already be discovered, we may wish our hypothesis generation procedure to generate a *new*, previously undiscovered hypothesis. We exclude this here because we could imagine the procedure being applied to $y - \phi'(x)$ where ϕ' is the already known factor that influences y.

replicability will follow. But if there are many plausible hypotheses, then this need not be the case. Third, crucially, note that *veracity* is not included as a criterion; that is, h need not necessarily be true. We do not require that $h \in \Phi^*$, nor do we require that changing $h(x)$ would change y.[76] Requiring veracity sets entirely too high a bar. Finally, both plausibility and comprehensibility can just as easily be defined as properties of the individual hypothesis generated by a procedure (we will apply these definitions below in our empirical work). In contrast, replicability can only be defined as a property of the overall procedure.

A.4 Human hypothesis generation

First, we consider human hypothesis generation, since this is a natural starting point and benchmark for more data-driven procedures such as the one we propose below. The psychology (and sociology) of the human hypothesis generation procedure is enormously complex (see for example Langley et al. (1987)), so our goal is only to capture a few key properties in a simple model. A few definitions will be helpful. Given a data set D, define a *matched pair* as two observations (y_0, x_0) and (y_1, x_1) from D, where $y_0 = 0$ and $y_1 = 1$. Let E be a set of matched pairs. A hypothesis h is *consistent* with a matched pair if $h(x_0) < h(x_1)$. We say that a hypothesis h is consistent with a set of pairs if it is consistent with all the pairs in the set. We write $\mathcal{H}_E$ for the set of hypotheses consistent with E, and we write $\mathcal{I}_E = \mathcal{I} \cap \mathcal{H}_E$ for the set of interpretable hypotheses consistent with E. With these definitions in hand, we can define the human hypothesis generation procedure.

Procedure. *Assume that we have some data set D. The human hypothesis-generating procedure is a creative process we denote by $\mathcal{P}_c$ and operates by:*

> **Cognitive Constraints***: Uniformly sampling (without replacement) n matched pairs to form a set E.*[77]
>
> **Inspiration***: Picking one hypothesis at random from $\mathcal{I}_E$, the set of interpretable hypotheses consistent with E.*
>
> **Advantageous prior***: Possibly picking true hypotheses with higher probability. With probability α, the hypothesis is chosen from Φ^*, the set of true hypotheses. With probability $1-\alpha$, the hypothesis is picked uniformly at random from $\mathcal{I}_E \setminus \Phi^*$, the set of untrue interpretable hypotheses consistent with E.*

The procedure $\mathcal{P}_c$ is parameterized by n (how much data people can meaningfully process) and α (the extent to which they have some special access to what is actually true).

That is, people look at a subset of data and within that subset look for something that differentiates positive from negative cases. Typically, they focus on differences that hold *in the dataset*, rather than "out of sample."

So how does human hypothesis generation fare on our three criteria? Almost by construction, it does well on comprehensibility. People produce hypotheses that make sense to us as people. But human hypothesis generation fares less well on our other two criteria.

[76] Specifically, if we changed only the part of x that affected $h(x)$ and held the rest constant, y should change. Importantly, note that veracity (unlike plausibility) does not depend on the data-generating process: it is a feature of the true function, not the specific way we draw data.

[77] Notice in our rendition, for simplicity, we have not allowed for intrinsic biases in what humans might notice, such as confirmatory or categorical biases that might lead them to systemically notice certain relationships that are not there. Such biases would worsen both human hypothesis generation and our suggested procedure.

Human discovery procedures are not particularly replicable. A large body of evidence shows that human judgments have a great deal of "noise": different people draw different conclusions from the same observations, and worse, the same person may notice different things at different times (Kahneman et al., 2022). More broadly, there is a great deal of randomness in what data are attended to and which hypotheses are inspired. This inherent noisiness of human judgments is embodied in our understanding of creativity. We do not just accommodate the lack of replicability; at times, we celebrate it: happy for the luck involved in the singular sparks of insight that advance our thinking.

Nor does the human procedure necessarily produce empirically plausible hypotheses. The reason is subtle, but important. To understand the nature of the problem, consider the following trivial stylized example. Suppose that $x = (x_1, \ldots, x_k)$ is a k-dimensional binary vector and that all k dimensions are comprehensible, so that the hypothesis $h_i(x) = x_i$ is in $\mathcal{I}$ for all i. Further, suppose that $f^*(x) = x_1$; that is, the true function relating x to y only depends on the first dimension of x. And to make matters simple, assume that $\mathcal{P}_{\mathcal{D}}(x)$ is uniform, meaning all possible values of x are equally likely. In this setting, the function h_1 is the only true hypothesis, and the only empirically plausible hypothesis. However, even in this stylized setup where the true hypothesis is actually quite simple, people can end up generating non-plausible hypotheses.

To see this, consider a pair of data points $(x_0, 0)$ and $(x_1, 1)$. Since p is uniform, x_0 and x_1 will differ on $\frac{k}{2}$ dimensions in expectation. So there are a number of interpretable, consistent, but implausible hypotheses. A person looking at only one pair of observations would have a high chance of generating an empirically implausible hypothesis. Of course, as the number of matched pairs n increases, the probability of discovering an implausible hypothesis declines. But the problem still remains. The intuition here is related to "over fitting": even though there is no noise in y, there is randomness in which observations happen to be in D, and even more so for the n pairs sampled in E. That randomness can lead to idiosyncratic differences between the $y = 0$ and $y = 1$ cases. As the number of comprehensible hypotheses gets large, there is a "curse of dimensionality": there are many plausible hypotheses for these idiosyncratic differences. That is, many different hypotheses can look good in sample, but they need not work out of sample. We realize that human-recognized patterns may not even be actually be present, which is why a first step in applied work is often to see if the hypothesis holds in a correlational sense.

A.5 Algorithmic hypothesis generation

We now consider how algorithms may help with hypothesis generation. We will assume, for simplicity, that for a given data set D, an algorithm m exists that can predict y from x. Specifically, we have access to a black box algorithm, which from any data set D produces $m(x)$ that predicts y out of sample. Given such a black box predictor, in principle, we already have one hypothesis-generating procedure: simply output $m(x)$. Since $m(x)$ predicts y from x, it is by construction empirically plausible. But in practice, $m(x)$ is highly unlikely to be comprehensible. Even simple machine learning algorithms rarely produce prediction functions that are meaningful hypotheses. This helps us see the strengths and weaknesses of both algorithms and humans for purposes of hypothesis generation:

- Human hypotheses are comprehensible but may not be empirically plausible.

- Algorithmic hypotheses are empirically plausible but may not be comprehensible.

Our goal is to marry people's unique knowledge of what is comprehensible with an algorithm's superior capacity to find meaningful correlations in data. One approach might be to formalize the set $\mathcal{I}$ and then focus on creating machine learning techniques that search over functions in $\mathcal{I}$. But mathematically characterizing $\mathcal{I}$ is often not possible. This is related to what Autor (2014) called "Polanyi's paradox," the idea that people's understanding of how the world works is largely beyond our capacity to explicitly describe it. Our failure to appreciate this paradox, and believe we understand more of our thinking than we do, is called the "introspection illusion" (Pronin, 2009). In our running example, how would we mathematically, or even verbally, characterize the set of functions of facial images (mugshots) that "make sense" to people? Instead we assume $\mathcal{I}$ remains unique, non-formalizable, tacit knowledge of people. Yet progress is still possible:

Procedure. *Suppose that $\mathcal{D}$ is some data-generating distribution, and that we have a data set D sampled from $\mathcal{D}$, and a density function p such that $p(x) > 0$ if and only if x can be sampled from $\mathcal{D}$. Further, assume that we have an algorithm m that predicts y, and some fixed values $\breve{m}$ and $\hat{m}$ such that*

$$\min_{\mathcal{D}}\{m(x)\} < \breve{m} < \hat{m} < \max_{\mathcal{D}}\{m(x)\}.$$

The algorithmic hypothesis procedure is a discovery process that operates by:

Morphing *Sample a random data point x_0 from the generative process p. Then, find points x^- and x^+ as solutions to the following problem:*

$$x^- = \arg\min_x\{\|x - x_0\| : m(x) \leq \breve{m} \text{ and } p(x) > 0\}$$

$$x^+ = \arg\min_x\{\|x - x_0\| : m(x) \geq \hat{m} \text{ and } p(x) > 0\}.$$

Naming *Show a human n such (x^-, x^+) pairs. Ask them to name a feature that differentiates these pairs. Specifically, they generate a hypothesis $h \in \mathcal{I}$, if there is one, that they view as differentiating the pairs.*

We write $\mathcal{P}_m$ for the algorithmic hypothesis procedure, or what we will also call the morphing hypothesis procedure.

The definitions of x^- and x^+ serve the same function as the matched pair in the human hypothesis procedure. However, thanks to the density function p, we can find x^- and x^+ such that the matching pair are both "close" (in some metric) to the original sampled point x_0. As we will see, this is the critical property of the algorithmic hypothesis procedure that improves the quality of the hypotheses that are output.

Morphing requires not just an algorithmic model of y but an algorithmic model of p. That model of p frees the algorithm from being constrained by the particular set of pairs found in the data set D and allows for the construction of entirely new data points that differ only along relevant dimensions, since we can use the prediction $m(x)$ to choose the new matching points. Put differently, for any given data point, this procedure allows us to construct new data points that answer the counterfactual question: How would this point be different if it had a higher or lower $m(x)$ value? Our approach to solving the given definitions to find x^- and x^+ is discussed in further detail in Section 5.

The second part of this procedure merely harnesses a *known human capacity*: the ability to notice differences in otherwise similar observations. Having people articulate hypotheses from looking at morphed data points rather than at raw data has two advantages:

- Left to their own devices, people seek to identify *any* differences across data points that differ in y. Because they only have one particular data set, that approach is prone to over fitting: differences that may hold in that particular sample but not others. But with our morphing procedure, humans are now looking for differences in $m(x)$, and we know $m(x)$ is a reliable out-of-sample predictor. By way of intuition, this is why we are usually better off interpreting a regression than individual data points that go into the regression. Although there is no noise in the setup we use here, the same basic intuition carries through.
- The matching of nearest neighbors reduces the curse of dimensionality. Pairs (x^-, x^+) will now have fewer plausible candidates for what may be different between them exactly because we have ensured they differ as little as possible.[78]

This procedure marries the algorithm's capacity to find signal with unique human knowledge of what is a meaningful hypothesis. Because people are not looking at actual data, they are effectively *naming* $m(x)$. They are projecting the algorithm into their own language—the set of hypotheses that are comprehensible. As a result, mechanically, this produces comprehensible hypotheses. At the same time, because $m(x)$ is known to have signal for y, this procedure is more likely to produce empirically plausible hypotheses.

We conclude by examining how our semi-automated procedure and human hypothesis generation compare. To do so, we will need to make an assumption about how the implausible hypotheses behave. In particular, we need to guarantee that in any given sample draw, the rate at which implausible hypotheses are consistent with any particular sample falls off sufficiently fast with sample size. To ensure this, we will assume the following two conditions hold for $\mathcal{D}$. First, for any distinct hypotheses $h, h' \in \mathcal{I}$ and any matching pair e, we assume hypotheses are consistent independently $\mathbb{P}_{\mathcal{D}}(h' \in \mathcal{I}_e \mid h \in \mathcal{I}_e) = \mathbb{P}_{\mathcal{D}}(h' \in \mathcal{I}_e)$. Second, for any implausible hypotheses $h, h' \in \mathcal{I} \setminus \Phi^*$ and any random pair of matching images (or random pair of matching morphs) e, we assume hypotheses are consistent equiprobably: $\mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_e) = \mathbb{P}_{\mathcal{D}}(h' \in \mathcal{I}_e)$. These assure the "concentration of mass" property we require (surely other assumptions would as well). Given these assumptions about $\mathcal{D}$, we can show the following.

Proposition 1. *Suppose that we have some data-generating distribution $\mathcal{D}$ with data set D. Let $\mathcal{P}_h$ be the human hypothesis procedure with n pairs and advantageous prior with parameter p. Let $\mathcal{P}_m$ be the algorithmic hypothesis procedure with n pairs and parameters $\check{m}$, $\hat{m}$. Assume also that the advantageous prior α is more advantageous than random choice; that is, assume that $\alpha > \frac{|\Phi^*|}{|\mathcal{I} \setminus \Phi^*|}$. There exist constants N, C_Φ and C_m such that if*

(i) People do not look at too many pairs $n < N$,

[78] An analogy with OLS provides an easy intuition for why this helps: A regression is in effect a way to calculate in-sample correlations, and the standard errors tell us whether those correlations are real or due to sampling noise. Adding controls lowers standard errors because lowering the residual noise makes it more likely that in-sample correlations are more likely to hold out-of-sample. That is the same effect here: By matching on $m(x)$ instead of y, we are reducing residual noise and increasing the chance that noticed patterns are genuine ones.

(ii) The problem is complex enough. That is, the relative number of comprehensible hypotheses that are true is sufficiently small: $\frac{|\Phi^*|}{|\mathcal{I}\setminus\Phi^*|} < C_\Phi$*, and*

(iii) Implausible hypotheses are consistent with matching morph pairs sufficiently rarely: $\mathbb{P}_\mathcal{D}(h \in \mathcal{I}_{e_m}) < C_m$ *for any hypothesis* $h \in \mathcal{I} \setminus \Phi^*$ *and any random matching morph pair* e_m*,*

then

- *Morphing produces plausible hypotheses at a higher rate:* $\pi(\mathcal{P}_m) \geq \pi(\mathcal{P}_c)$*,*
- *Morphing is more replicable:* $\rho(\mathcal{P}_m) \geq \rho(\mathcal{P}_c)$*, and*
- *Both procedures produce comprehensible hypotheses* $\xi(\mathcal{P}_m) = \xi(\mathcal{P}_c) = 1$*.*

(See below for a sketch proof of Proposition 1.)

The proposition highlights how a morphing procedure could be useful in principle. We now assess whether it works in practice. In what follows, we will implement this procedure to generate a hypothesis $h(x)$. In the ideal setup of our framework, such an h represents meaningful communication. Whether that is actually the case is an empirical question. We do not test for comprehensibility because, as pointed out, by definition any h produced by people is comprehensible. But we will directly measure whether what people see is also what the algorithm "sees": whether in fact $h(x)$ predicts $m(x)$. And we will assess the procedure on two dimensions we have already described: reliability (what fraction of subjects name the same h); and plausibility (whether h predicts y).

Finally, it is worth noting a few additional advantages of algorithmic generation that are not highlighted by this proposition. First, though it is not explicit here, given a set of *known* hypotheses, we can orthogonalize with respect to those dimensions to ensure that the algorithm is producing something novel.[79] Second, other methods of producing hypotheses (observation, conversation, introspection) may produce theories that are hard to measure in data. By construction, our procedure only produces hypotheses that are measurable.

A.6 A proof of Proposition 1

Sketch proof of Proposition 1. Suppose that we have some data generating distribution $\mathcal{D}$. Let D be a data set drawn from $\mathcal{D}$. Let $\mathcal{P}_c$ be the human hypothesis procedure with n pairs and advantageous prior with parameter α. Let $\mathcal{P}_m$ be the algorithmic hypothesis procedure with n pairs and parameters $\check{m}$, $\hat{m}$. Let E_c be a random set of n matched pairs randomly drawn from a matching process on D, and let E_m be a random set of n matched morph pairs randomly drawn from the morphing process.

Our strategy for the proof will be as follows. First, we will use the assumed conditions on $\mathcal{P}_c$ and $\mathcal{P}_m$ to derive some more convenient identities used throughout the proof. We will then find some upper and lower bounds on $\mathbb{P}_\mathcal{D}(\mathcal{P}(D) = h)$ for both hypothesis generating procedures, and under different conditions on $h \in \mathcal{I}$. Finally, given these bounds, we will use the definitions of plausibility and reproducibility to directly prove the claims from the proposition statement.

We begin by producing some convenient identities involving the input parameters. First, we let $K = |\mathcal{I} \setminus \Phi^*|$, and assume that we have some fixed positive integer $N > n$. Next, since

[79] In our particular application, we do not do this, in part because we are curious to explicitly examine algorithms' capacity to rediscover known hypotheses.

we have assumed that implausible hypotheses are consistent equiprobably, we know that $\mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_{e_c})$ and $\mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_{e_m})$ are both independent of $h \in \mathcal{I} \setminus \Phi^*$, $e_c \in E_c$, and $e_m \in E_m$. Hence, we can define constants ξ_c and ξ_m by

$$\xi_c = \mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_{e_c}) \quad \text{and} \quad \xi_m = \mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_{e_m}),$$

where $h \in \mathcal{I} \setminus \Phi^*$ is any implausible hypothesis, $e_c \in E_c$ is any single matching pair, and $e_m \in E_m$ is any single matched morph pair. Further, since matching pairs are sampled independently,

$$\mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_{E_c}) = \xi_c^n, \quad \text{and} \quad \mathbb{P}_{\mathcal{D}}(h \in \mathcal{I}_{E_m}) = \xi_m^n$$

whenever $h \in \mathcal{I} \setminus \Phi^*$ is an implausible hypothesis. We now define constants $\beta_h = \xi_h^n$ and $\beta_m = \xi_m^n$. Now, the given assumptions on $\mathcal{I}$ and Φ imply that we can fix C_Φ so that

$$C_\Phi \leq \xi_h^N,$$

which implies that

$$\beta_h = \xi_h^n \geq \xi_h^N \geq \frac{|\Phi^*|}{K}. \tag{2}$$

Similarly, the given assumption on consistent hypotheses implies that we can fix C_m such that

$$C_m \leq \left(1 - \alpha^{\frac{1}{K}}\right)^{\frac{1}{N}},$$

which implies that

$$\beta_m = \xi_m^n \leq \xi_m^N = C_m^N \leq 1 - \alpha^{\frac{1}{K}},$$

and hence that

$$\alpha < (1 - \beta_m)^K. \tag{3}$$

We now turn our attention to proving some identities involving the hypotheses in $\mathcal{H}$. We have assumed that for a given matched pair e, hypotheses in $\mathcal{H}$ are consistent with e independently. For any set of hypotheses $\mathcal{I}_E$ such that $\Phi^* \subseteq \mathcal{I}_E \subseteq \mathcal{I}$, this implies that

$$\mathbb{P}_{\mathcal{D}}(\mathcal{I}_{E_c} = \mathcal{I}_E) = \beta_c^k (1 - \beta_c)^{K-k}, \quad \text{and} \quad \mathbb{P}_{\mathcal{D}}(\mathcal{I}_{E_m} = \mathcal{I}_E) = \beta_m^k (1 - \beta_m)^{K-k}. \tag{4}$$

We will also use the fact that there are exactly $\binom{K}{k}$ distinct possible values for $\mathcal{I}_E$ such that $|\mathcal{I}_E| = |\Phi^*| + k$ and $\Phi^* \subseteq \mathcal{I}_E \subseteq \mathcal{I}$.

We will now find bounds on the probability that a given hypothesis h is an element of $\mathcal{I}_E$, for both the human discovery process and the morph discovery process. We know that for any hypothesis $h \in \mathcal{H}$,

$$\mathbb{P}_{\mathcal{D}}(\mathcal{P}_c(D) = h \mid \mathcal{I}_{E_c}) = \begin{cases} 0 & \text{if } h \notin \mathcal{I}_{E_c} \\ \frac{1-\alpha}{|\mathcal{I}_{E_c}|} & \text{if } h \in \mathcal{I}_{E_c} \setminus \Phi^* \\ \frac{\alpha}{|\Phi^*|} & \text{if } h \in \Phi^*. \end{cases} \tag{5}$$

For plausible hypotheses, (5) is sufficient to calculate an unconditional likelihood that $\mathcal{P}_c$ produces a plausible hypothesis as output. That is, for any $h \in \Phi^*$,

$$\mathbb{P}_{\mathcal{D}}(\mathcal{P}_c(D) = h \mid h \in \Phi^*) = \frac{\alpha}{|\Phi^*|}. \tag{6}$$

Conversely, for the morphing discovery process, we see that

$$\mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid \mathcal{I}_{E_m}) = \begin{cases} 0 & \text{if } h \notin \mathcal{I}_{E_m} \\ \frac{1}{|\mathcal{I}_{E_m}|} & \text{if } h \in \mathcal{I}_{E_m}. \end{cases}$$

Focusing on the probability of producing an implausible hypothesis, we see that

$$\begin{aligned} \mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid h \in \mathcal{I} \setminus \Phi^*) &= \sum_{E_m \subseteq \mathcal{I}} \mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid \mathcal{I}_{E_m}) \cdot \mathbb{P}_{\mathcal{D}}(E_m) \\ &= \sum_{E_h \subseteq \mathcal{I}} \left[\frac{1}{|\mathcal{I}_{E_m}|}\right] \cdot \mathbb{1}\{h \in E_h\} \cdot \mathbb{P}_{\mathcal{D}}(E_m) \\ &= \sum_{k=0}^{K-1} \frac{1}{k+1+|\Phi^*|} \cdot \binom{K-1}{k} \cdot \beta_m^{k+1}(1-\beta_m)^{(K-1)-k} \\ &\leq \frac{1}{K} \cdot \sum_{k=0}^{K-1} \cdot \binom{K}{k+1} \cdot \beta_m^{k+1}(1-\beta_m)^{K-(k+1)} \\ &= \frac{1-(1-\beta_m)^K}{K}. \end{aligned}$$

In order to form the inequality above, we have used the identity that for any positive integer k such that $k \leq K$,

$$\frac{1}{k+1}\binom{K}{k} = \frac{1}{K+1}\binom{K+1}{k+1}.$$

Using complementarity and the assumption of equiprobability, we know that

$$\mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid h \in \Phi^*) = \frac{1}{|\Phi^*|} \cdot (1 - K \cdot \mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid h \in \mathcal{I} \setminus \Phi^*)).$$

Hence,

$$\mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid h \in \Phi^*) \geq \frac{(1-\beta_m)^K}{|\Phi^*|}. \tag{7}$$

Having now derived various upper and lower bounds for the probability that a given hypothesis is returned by either discovery process, we are now ready to proceed with the body of the proof. We will show by direct substitution that each of the claims made in the proposition hold.

We will begin with plausibility. We know by definition that a hypothesis $h \in \mathcal{I}$ is plausible if and only if $h \in \Phi^*$. Hence, we can measure the plausibility of either procedure $\mathcal{P}$ by

$$\pi(\mathcal{P}) = \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D) \in \Phi^*) = \sum_{h \in \Phi^*} \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D) = h).$$

But for any plausible hypothesis $h \in \Phi^*$, (7) and (6) mean that

$$\mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid h \in \Phi^*) - \mathbb{P}_{\mathcal{D}}(\mathcal{P}_c(D) = h \mid h \in \Phi^*) \geq \frac{(1-\beta_m)^K}{|\Phi^*|} - \frac{\alpha}{|\Phi^*|} \geq 0, \tag{8}$$

using the assumption that $\alpha < C_\alpha$ and (3). But this directly implies that

$$\pi(\mathcal{P}_m) \geq \pi(\mathcal{P}_c),$$

which proves the first claim of the proposition.

We will now focus on the replicability of $\mathcal{P}_c$ and $\mathcal{P}_m$. Let D' be another draw from the data-generating process $\mathcal{D}$. Then for either generating procedure $\mathcal{P}$ we see that

$$\rho(\mathcal{P}) = \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D) = \mathcal{P}(D')) = \sum_{h \in \mathcal{I}} \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D) = h) \cdot \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D') = h).$$

Hence,

$$\rho(\mathcal{P}) = |\Phi^*| \cdot \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D) = h \mid h \in \Phi^*)^2 + K \cdot \mathbb{P}_{\mathcal{D}}(\mathcal{P}(D) = h \mid h \in \mathcal{I} \setminus \Phi^*)^2. \tag{9}$$

Now, using (9), we consider the difference in replicability between the human and algorithmic hypothesis generating procedures directly. For convenience, we define the variable

$$\gamma_m = \mathbb{P}_{\mathcal{D}}(\mathcal{P}_m(D) = h \mid h \in \Phi^*).$$

Then (9) becomes

$$\rho(\mathcal{P}_m) = |\Phi^*| \cdot \gamma_m^2 + K \cdot \left(\frac{1 - |\Phi^*| \cdot \gamma_m}{K}\right)^2 = |\Phi^*| \cdot \gamma_m^2 + \frac{(1 - |\Phi^*| \cdot \gamma_m)^2}{K}.$$

Then we can use (9) to see that

$$\begin{aligned}
\rho(\mathcal{P}_m) - \rho(\mathcal{P}_c) &= |\Phi^*| \cdot \left(\gamma_m^2 - \frac{\alpha^2}{|\Phi^*|^2}\right) + \frac{1}{K} \cdot \left((1 - |\Phi^*| \cdot \gamma_m)^2 - (1 - \alpha)^2\right) \\
&= |\Phi^*| \cdot \left(\gamma_m^2 - \frac{\alpha^2}{|\Phi^*|^2}\right) - \frac{2|\Phi^*|}{K} \cdot \left(\gamma_m - \frac{\alpha}{|\Phi^*|}\right) + \frac{|\Phi^*|^2}{K}\left(\gamma_m^2 - \frac{\alpha^2}{|\Phi^*|^2}\right) \\
&= |\Phi^*| \cdot \left(\gamma_m - \frac{\alpha}{|\Phi^*|}\right) \cdot \left[\left(1 + \frac{|\Phi^*|}{K}\right) \cdot \left(\gamma_m + \frac{\alpha}{|\Phi^*|}\right) - \frac{2}{K}\right].
\end{aligned}$$

Now, clearly $|\Phi^*|$ is non-negative. Further, (8) and the assumption of equiprobability implies that $\gamma_m \geq \frac{\alpha}{|\Phi^*|}$, so the second factor of the above expression is also non-negative. For the third and final term, we re-use the result that $\gamma_m \geq \frac{\alpha}{|\Phi^*|}$ and the assumption that $\alpha > \frac{|\Phi^*|}{K}$ to see that

$$\left(1 + \frac{|\Phi^*|}{K}\right) \cdot \left(\gamma_m + \frac{\alpha}{|\Phi^*|}\right) > \left(1 + \frac{|\Phi^*|}{K}\right) \cdot \frac{2}{K} > \frac{2}{K}.$$

Thus, the third term of the product above is also strictly positive. But this directly implies that

$$\rho(\mathcal{P}_m) \geq \rho(\mathcal{P}_c),$$

which proves the second claim of the proposition, and completes the proof. □

B Appendix B: Data and Institutional Details

B.1 Pre-trial detention decisions

When someone is arrested in the United States, they must be brought in front of a judge (usually within 24–28 hours) to decide what should happen to the defendant as they await resolution of their case. This decision under the law is supposed to hinge on the defendant's risk of flight (skipping future court hearings) or public safety risk (re-arrest). That is, it is supposed to hinge on a *prediction*. In most jurisdictions, the decision options available to the judge at this hearing include:

- Release the defendant outright, often known as released on recognizance (ROR),
- Release the defendant conditional on their providing some collateral, such as cash bail, with the intention of ensuring re-appearance at future court dates,
- Release the defendant with the requirement that they be monitored by some electronic location tracking device,
- Order the defendant detained.

One implication is that defendants can wind up in jail awaiting trial for at least two reasons, first because the judge explicitly ordered them to jail, and second because the defendant cannot come up with the required collateral for release. While judges are supposed to set collateral requirements that defendants can come up with to get released, in practice (from our own observations of court proceedings in different jurisdictions) it would appear that judges sometimes intentionally set bail at a level that the defendant *cannot* make, as a sort of back-door way to ensure detention. In our own analysis, we follow Kleinberg et al. (2018) and abstract from the nuances of this range of choices and just focus on the binary outcome of whether the defendant was detained (either because they were remanded by the judge outright, or had a cash bail set above what they could pay) versus were released (regardless of whether they were ROR'd or assigned a bail they were able to post).

This process can vary somewhat across different jurisdictions within the US. For example, in some places, judges do not have the option of explicitly ordering a defendant sent to jail without the possibility of posting collateral for release. (That is, the judge cannot order detention directly.) Some jurisdictions allow judges to release defendants under an order to participate in pre-trial services, which can include periodic reporting to a pre-trial services officer. Some jurisdictions are beginning to prohibit judges from requiring they post collateral or bail to get released, either just for selected offenses or for all cases across the board. Some jurisdictions require judges to consider only flight risk, not safety risk.

In the specific jurisdiction from which we have obtained data here, Mecklenburg County, North Carolina, the very first hearing for the defendant is overseen not by a judge, but by a "magistrate" (who is like a judge, but is not elected). Defendants not released by the magistrate are booked into jail and see a judge the next day (Redcross et al., 2019). Starting in 2014, judges were given access to a pre-trial risk prediction tool developed by the Arnold Foundation called the Public Safety Assessment (Redcross et al., 2019). The PSA gives judges predictions from a logistic regression for three separate outcomes: (1) risk of failure to appear (FTA) in court at a required future hearing; (2) risk of any new criminal activity (NCA); and (3) risk of any new violent criminal activity (NVCA). The PSA makes these predictions using factors like age, current charge, and prior record.[80] Because defendants can only be detained if the magistrate and judge agree on detention, and because the magistrate's decision is made in the shadow of the judge, and because (more pragmatically) the data we have do not separately identify the magistrate's decision from that of the judge, we follow Redcross et al. (2019) and combine both decisions into a single detain-versus-release outcome.

How do these cases get resolved? A large share will simply wind up being dropped (see for example Agan et al. (2021)). Among those cases that result in a finding of guilt, the large majority will be resolved through a plea deal rather than through a trial. The decision about what the punishment should be for a guilty defendant depends on a wider range of factors

[80] See `https://advancingpretrial.org/psa/factors/`.

than does the pre-trial detention decision. Beyond recidivism risk (key for pre-trial detention decisions), sentencing decisions also depend on considerations such as society's sense of just desserts, the defendant's remorse, and impacts on victims.

B.2 Mecklenburg County criminal justice data

We downloaded a total of 81,166 arrest records from the public MCSO website. We apply a number of filters to these data to form our final analysis data sets that exclude cases that are missing some key information needed for our analysis, contain some obvious data error, or capture cases that are not subject to a normal pre-trial detention decision by the judge. The complete list of filters are described in Table A.A.I and include:

- We drop cases that are missing at least one piece of key information, such as the defendant's mugshot (a key input to predicting judge decisions), the court case ID (which we need to link the criminal justice data sets together), the charge for which the defendant was arrested (which we need to predict defendant re-arrest risk), and bond information or jail stay information (which is part of determining whether defendants are detained versus released).
- The case is listed as a "non-arrest," which often means this is related to a probation or parole violation or a case related to a federal warrant. We exclude these because the pre-trial detention decisions are typically quite different from "normal" cases.
- There is clearly some error in the data, for example, the arrest date is listed as coming after the date the case was resolved in court.
- The arrest was disposed of within three days. These are excluded since the magistrate or judge decision may be quite different in these cases; that is, if the strength or weakness of the case is observable to magistrates and judges, they might automatically release the case if they realize it will just be dropped very quickly.

The filters taken together eliminate about one-third of the arrests that occurred during our observation period.

We also apply one final filter to the lock-box hold-out data set as well. Part of this hold-out data set consists of arrests made in the last 6 months of our data period, so that we can test the predictive accuracy of our models in a new time period. To avoid inadvertent information leakage, we drop cases for people who were arrested during this time period and also show up as having been arrested in the training data set.

To construct our measure of "release," we count everyone who left jail not more than three days after arrest. This will include everyone who was released on their own recognizance (RORd) by the judge, as well as people who are assigned cash bail by the judge (they are required to post collateral to get released) and are able to make that bail fairly quickly. In the data, we see only a modest share of people get released much more than three days after the date of the arrest, so our results should not be very sensitive to adjusting this threshold out further.

Our measure of "re-arrest" combines information from the MCSO data on all arrests, together with the NCAOC data set on when each case (past arrest) gets resolved. So for a given arrest, we can see whether the defendant has a new arrest that shows up in the MCSO data set that is filed prior to resolution of the initial arrest according to the NCAOC data.

Unfortunately, our data do not allow us to construct a usable measure of whether the

defendant skips court (or "failure to appear," FTA). In principle, that could create an omitted variable bias concern, if the defendant characteristics we examine in this paper were correlated with FTA. But since the defendant characteristics are facial features, we think this risk of bias (in the econometric sense of the term) is not serious.

From the raw data we construct features corresponding to:

- The type of charge for which the defendant has been arrested (violent crimes, property crimes, drug crimes, or other offenses), and
- Detailed measures of whether the defendant has been convicted of these different types of crimes at different points in the past 1, 3, 5 and 10 years.

In nearly half of all arrests, the defendant is charged with more than one offense. We follow the usual approach within criminology and classify each case by the most serious charge using the FBI's Uniform Crime Reporting system hierarchy. We then group crimes into our four broad categories of crime types (violent, property, drug and other), combining arrests for both more and less serious versions of each type of crime in each category. (So, for example, assaults that fall into the FBI "part 1" or more serious category would get counted as violent crimes alongside assaults that are counted as "part 2" crimes.) For predicting defendant risk, we also experiment with providing the algorithm access to more detailed current charge descriptions (like "possession of less than 0.5 ounces of marijuana," "larceny" and "armed robbery") as well as higher-level aggregations of charges (drug, property or violent crime charges).

Because the MCSO's website makes arrest data (and hence mugshots) available for the past 3 years on a rolling basis, other researchers can use the code we post to scrape mugshots off the MCSO website and carry out a similar analysis to what is reported here.

The mugshot photos are taken from a standard distance with the defendant standing in front of a flat gray wall looking at the camera. There are no side-view facial images in this dataset. Defendants are presumably asked to remove glasses or hats, since none of the images include those accoutrements. It is usually possible to see part of the defendant's shirt. Most defendants are wearing whatever they had on when they were arrested, although some defendants look to be wearing jumpsuits of the sort that many correctional facilities issue to inmates. These may be defendants who were charged with an offense they allegedly committed while in detention, or with an offense they allegedly committed prior to being detained but where sufficient evidence for charging was not possible to accumulate until after the defendant was already detained for some other offense.

C Appendix C: Methods

In this appendix, we discuss our methods for predicting judges' decisions and defendant risk, generating mugshots using GANs, and our procedure for generating morphed image pairs, including how we iterate our procedure and orthogonalize subsequent image morphings for the hypotheses discovered during earlier morphing cycles.

C.1 Predicting judges' decisions and defendant risk

The data we have downloaded from North Carolina include both structured variables (age, current charge, etc.) and unstructured, high-dimensional data sources like mugshot images.

As noted in the text, we build separate types of models for the structured data (gradient boosted decision trees) and unstructured data (convolutional neural networks, or CNNs). For our models that rely on both structured and unstructured data, we use a stacking procedure that forms new predictions that are weighted averages of the structured data predictor and unstructured data predictor, with the data used to select the weight. Since we are using standard machine learning methods at this stage of our analysis, we focus our discussion here on high-level descriptions.[81]

A decision tree recursively partitions the data through a series of top-down "splits" of the data by values of the features, x, where each split is selected to minimize some loss function $L(y, m(x))$ (for example, likelihood for binary outcomes or squared error for continuous outcomes). The result is a tree with M terminal nodes, where each terminal node is internally as similar as possible with respect to y. If each node i covers a region of the feature space R_i, then the prediction within each node is $c_i = \mathbb{P}(y = 1|x \in R_i)$, and the prediction from this decision tree is given by

$$m_s(x) = \sum_{i=1}^{M} c_i \cdot \mathbb{1}\{x \in R_i\},$$

where $\mathbb{1}$ is the indicator function, which is 1 if the argument is true, and zero otherwise. The "deeper" the tree (the more levels of splits), the better the tree is at fitting the relationship between x and y, but the more unstable (sensitive to small changes in the data) the tree can be. This challenge is often overcome by generating multiple versions of the predictor by perturbing either the training data set or the algorithm construction method and then combining them, what Breiman (1998) calls "perturbing and combining." A different approach (the one we use here) is to build a series of "shallower" trees that are less unstable, but at the cost of fitting the data less well than a deeper tree would. To reduce bias in the statistical sense of the term, we use boosting to build a series of trees iteratively, which increasingly up-weight the observations most poorly predicted to that point.

The logic behind the CNN method is perhaps easiest to see by considering its alternatives. To an algorithm, a 512×512 black-and-white image is essentially just $262,144$ pixel values.[82] It is clear that a simple linear function would be of little use, since the meaning of any one pixel's shading depends on other pixels. But estimating a regression that tried to allow every one of the $262,144$ pixel values to interact with every other pixel becomes intractable. This approach would also ignore the topography of the data; in an image, the shading of a pixel will be correlated with that of nearby pixels. This helps us see why early AI attempts to go directly from the "raw" image to prediction led to poor performance.

The basic idea behind a deep-learning neural network is to construct a series of intermediate layers between the inputs and the final classification outputs where the earliest layers try to learn the most concrete concepts (for images this would be, for example, edges or corners), and each subsequent layer learns increasingly abstract, complicated concepts (such

[81] For excellent overviews of decision trees and gradient boosting methods at various levels of technical detail, see for example Bishop and Nasrabadi (2006), Breiman (2001), Breiman et al. (2017), Freund et al. (1999), Hastie et al. (2009), and James et al. (2013). Examples of excellent discussions of deep-learning methods at various levels of technical complexity include Yegnanarayana (2009), LeCun et al. (2010), Krizhevsky et al. (2012), LeCun et al. (2015), Nielsen (2015), Rawat and Wang (2017), and Gurney (2018).

[82] For a color image, there are three times as many values, since pixels have red, blue and green shadings.

as what combination of edges, corners, etc. make up an eye, and then what combination of eye-like, nose-like and mouth-like concepts, in what relation to one another, make up a face, etc.). Because some of the early intermediate features are not specific to any given image application, it is possible to improve a CNN's performance through "pre-training" and learning some of these intermediate concepts from other data sources. A convolutional neural network (CNN) is a specific version of a neural network designed to work particularly well with image processing tasks. The specific version of a CNN that we estimate here is known as a residual network, which enables the estimation of more accurate deeper networks; see He et al. (2016).

The main binary outcome variable (y) we seek to predict in this classification exercise is an indicator for whether the judge detains rather than releases a given defendant as they await resolution of their case. For purposes of being able to morph faces with our generative adversarial network (GAN) for basic demographic features, we estimate a "multi-head" CNN that predicts four outcomes simultaneously:

- Release (released versus detained),
- Gender (male versus female),
- Race (Black versus white or other race),
- Age (above or below the sample median age of 29).

As noted above, what slightly complicates our analysis here is the fact that our "inputs" to predicting the judges' decision (x) include both image data (the red, green and blue shading values for each pixel in the images) and standard structured variables. Estimating a single residual network using both types of data creates estimation challenges because the network can "learn" the signal in the structured data much more easily than it can from the image data, and so winds up under-optimizing the available signal from the images. To address that problem, we estimate the stacked ensemble algorithm described in the main text and above.

The image data are fed into a 50-layer residual network ("resnet50") that consists of 4 convolutional blocks and 2048 output neurons, using a gentle decay learning rate schedule (see He et al. (2016)). Because the more basic features of images are not specific to the types of images being analyzed, we can improve performance of this network by pre-training it on a separate set of images. The resnet50 we use here was pre-trained on ImageNet data[83] with an ACC@1 score of 76.130 and ACC@5 of 92.862. We also tried a 15-layer residual network, or 'resnet15,' and a Mobile Net V2, and selected the resnet50 as best given its out-of-sample predictive accuracy.

To estimate defendant risk of re-arrest, we use *only* the sample of defendants who are released by the judge as our training data set. The reason is that re-arrest is defined as having a new arrest in between the original or focal arrest and resolution of that case (dismissal, a finding of innocence or guilt, etc.), since the judge's release decision is supposed to hinge on risk of re-arrest through case resolution. Defendants who are detained through the end of their case are missing data on whether they would have been re-arrested had they been released. Using this subsample as our training data set, we build a gradient boosted tree algorithm whose inputs are the structured data we have from Mecklenburg County. Specifically, we give the algorithm access to detailed current charge information (we partition 824 unique charge

[83] `https://www.image-net.org/`

descriptions into four categories: violent, drug, property, and gun-crime charges) prior record information, and demographic variables. The AUC of this algorithm in the validation set of released defendants equals 0.735, which is comparable to other risk predictors such as the proof-of-concept model built using New York City data in Kleinberg et al. (2018), which had an AUC of 0.707 in predicting FTA risk (the outcome judges are asked to consider in New York State). For purposes of the analysis presented in the main exhibits, we can assign predicted re-arrest risk values to everyone in the validation data set (since that prediction is a function of structured covariates available for everyone) that enables us to, for example, regress detention outcomes against predicted risk and other variables.

C.2 Alternative methods for algorithmic interpretability

The problem we face—understanding what our algorithm sees in the face—has emerged as a central challenge in machine learning research. A variety of techniques have been developed for interpreting or explaining how machine learning algorithms form their predictions (see Marcinkevičs and Vogt (2020) for a recent review).[84] Here, we give a high-level overview of how those techniques relate to our work.

A first major divide in the literature is whether we are seeking explanations that are already measured. One category of explanability methods can only provide explanations using measured high-level features. For example, Li et al. (2018), Ghorbani et al. (2019), Zhang et al. (2018), and Chen et al. (2019) among others develop interpretability tools that highlight not individual pixels that are important for classification, as in saliency maps, but higher-level *concepts* or prototypical parts within these images, such as wheels helping classify the presence of a van in an image. But all these approaches require the explanatory features to already be coded: the data must contain for each image, for example, information on whether "wheel" was present or not.[85] In these examples, the goal is typically not discovery but instead either to explain the model to people to aid in decisions, sometimes as required by explanation (Wachter et al., 2018), or to assess the robustness of models, such as whether a breast cancer detector is looking in the right place (Bai et al., 2021). Moreover, since the potential explanations are already in the data set, one could go further: rather than building a black-box model and explaining it, build one that is explainable to begin with.[86] All these techniques can be used for unstructured data, such as images or text, but only when the potential explanations are already coded in the data.

In the same category, closer to our approach is work on controllable generation (Lee and Seok, 2019). This work also relies on an unsupervised model (often a GAN), but the goal here is to be able to generate images with certain characteristics, which are once again the features already measured in the data. For example, rather than generating synthetic faces,

[84] For simplicity, we will use the phrase "explanations" to describe what we seek from the model. In the literature, some use the phrase "explanations" and "interpretations" differently.

[85] In our example, our mugshots do not begin with any annotations. Moreover, if we were to choose what to annotate, we would choose the features we already believe are important, such as competence and trustworthiness. The discovered features (e.g., "heavy-faced" or "well-groomed") were discovered from the pixels not because we had already chosen to measure them. We annotate them in the data once they have been discovered as part of the validation exercise.

[86] See, for example, Holte (1993), Rudin et al. (2010), Freitas (2014), Letham et al. (2015), Angelino et al. (2018), Jung et al. (2017), Chen and Rudin (2018), Ustun and Rudin (2019), Rudin (2019), and the references therein.

the goal would be to generate an old face, and this is done when age is measured in the data during training.[87]

By way of contrast, our data do not already have "heavy-faced" or "well-groomed" defined. Without these annotations, the previous methods cannot work. To make them work, one could imagine collecting labels on an extremely large set of facial features and then apply one of the approaches described above. The challenge in doing this is the enormous effort needed to codify so many different facial features.[88] In some sense, it is akin to the problem of hypothesis generation: what features should we annotate?

More recent work on interpretability has focused on situations where the potential explanation is not already coded in the data (some of it referred to as "counterfactual explanations"). Here, the idea is to morph input images, as we are doing, rather than simply highlight regions. We are far from the first to combine the idea of a generative model with a predictive model to provide explanations (Chang et al., 2018). In a different context, Miller et al. (2019) introduces an idea much like our procedure, where a Variational Autoencoder is used as the generative model. More recent work in this same vein can be found in Ghandeharioun et al. (2021); Lang et al. (2021) and Liu et al. (2019). Our approach firmly fits in this last category of approaches. Some of these recent attempts to generate counterfactual images use an approach that trains the GAN and the predictor $m(x)$ together at the same time (Lang et al., 2021; Ghandeharioun et al., 2021). In principle, it is possible these alternative methods could produce even better counter-factual morphings than does our own procedure, although given the quality of our own morphs there would seem to be at best modest room for improvement. In any case our own procedure has the practical advantage of requiring substantially less computational time to implement.

C.3 Generative Adversarial Networks

Generative adversarial networks (GANs) were developed initially as procedures for creating realistic, but fake, images (see for example Goodfellow et al. (2014b), Goodfellow et al. (2020)).

As noted in the text, a GAN is built by training two algorithms that "compete" with each other, the *generator* G and the *classifier* C: the generator creates synthetic images and the classifier (or "discriminator"), presented with synthetic or real images, tries to distinguish

[87] One could think of our approach, in spirit, as controllable generation but for situations where rather than generating for a known feature (e.g., age), we are generating according to a predictor (e.g., predicted detention probability). While conceptually these are the same, in implementation, we take a slightly different approach. Typically, for controllable generation, the GAN itself is trained differently so that individual dimensions of interest (e.g., age) are represented individually in the latent space. We instead built a generic mugshot GAN and morph. The reason we chose that approach is that, unlike age, the prediction of detention itself is a very "noisy" label, an imperfect judgment of detention risk. So while the differences between faces in age is quite dramatic, the differences in detention probability can be more subtle.

[88] A recent ambitious paper has tried to tackle this problem. Peterson et al. (2022) collected millions of labels on hundreds of facial features and then created a predictive model of them for synthetic faces. The challenge, however, is that this model is built on synthetic faces, whereas we would need such a model for actual images (mugshots). Deep learning models are known to not transfer across distributions. In fact, when we attempt to use the results of this paper, we find our mugshots do not map into these synthetic faces in any meaningful way. The failure is a reminder that while humans tend to think of "faces" in the abstract, algorithms model very specific distributions of pixel combinations. It is why we must build our own generative model of mugshots rather than use extremely well-developed generative models of "faces."

which is which. A good discriminator pressures the generator to produce images that are harder to distinguish from real, and in turn, a good generator pressures the classifier to get better at discriminating real from synthetic images. Data on actual faces is used to train the discriminator, which then results in the generator being trained as it seeks to fool the discriminator.

Specifically, the generator is a function that maps a (typically multivariate) random variable z to the target space of images in $\mathbb{R}^k$. That is, the generator produces random images $G(z)$ that seek to follow the distribution of the actual data set of real images, $p(x)$. The discriminator outputs the probability a given image x is a real image, $C(x) \in [0, 1]$, seeking to maximize this probability for real images and minimizing the probability for generated images $G(z)$. The loss function for C given generator G equals:

$$L^C = -E_{x\sim p(x)}[\log C(x)] - E_{z\sim p_z}[\log(1 - C(G(z)))].$$

The generator seeks to increase the chances the discriminator *incorrectly* classifies generated images as real images, or $C(G(z))$. The loss function for the generator is $E_{z\sim p_z}[\log(1 - C(G(z)))]$, although in other applications variations of this function are often used instead. The two algorithms essentially "play" against one other trying to create fake images that pass as real ones, and detect which images are fake. The objective function for the GAN is:

$$\min_G \max_D E_{x\sim p(x)}[\log C(x)] + E_{z\sim p_z}[\log(1 - C(G(z)))].$$

With machine learning, the performance of both C and G improve with successive iterations of training. A perfect G would output images where the classifier C does no better than random guessing. Such a generator would by definition limit itself to the same input space that defines real images; that is, the manifold of faces.

We use a StyleGAN2 developed by Karras et al. (2019), which is widely regarded as one of the most successful GAN architectures to date. Our GAN is trained on 33,100 mugshot images, each of which is structured as 512 pixels by 512 pixels, with a black boundary and centered faces.

One common measure for assessing a GAN's quality is the Frechet inception distance (FID) (Heusel et al., 2017), which is a measure of the difference between the distribution of GAN-generated images relative to the original images used to train the GAN.[89] On our subsample of male arrestees in the Mecklenburg data set, we obtain an FID of 1.71. By way of comparison, StyleGAN2 trained on the flicker-faces HQ data set (FFHQ), which contains 70,000 high-quality, high-resolution (1024x1024) images, equals 2.84.[90] We likely do better because the space of mugshots is a smaller, less rich space than the space of faces in the Flickr dataset.

Another pair of performance measures we use are precision and recall (Sajjadi et al., 2018), which are analogous to, but distinct from, common metrics of the same name used in

[89]Calculation of the FID measure begins with a general off-the-shelf image CNN (an Inception V3 classifier) and then uses the final layer of that classifier as a way to represent images. We then calculate the distribution of real and synthetic images in this representation space. The FID metric is the square of the Wasserstein distance between these two distributions, with lower values indicating better performance.

[90]As noted above, to avoid stereotyping in discussions of crime and criminal justice, we illustrate the key ideas in our paper using images just for non-Hispanic white males. So the GAN performance statistics we report here are from a StyleGAN2 trained just on males in our mugshot data set, which as shown in Table I accounts for the large majority of our sample.

predictive modeling. Precision measures the chance that a randomly generated image from the GAN is close to some real image from the training data, while recall measures the chance that a random image from the training data is close to some image generated by the GAN. Or, roughly speaking, precision is how often images with a positive $\widehat{p}(x)$ look like a face, while recall measures how much of the training data is assigned a positive $\widehat{p}(x)$ by the GAN. Our GAN has a precision of 0.7784 and a recall of 0.5741; by comparison, a StyleGAN or StyleGAN2 trained on the FFHQ dataset can achieve a precision up to 0.721 and a recall of 0.492 (Karras et al., 2020) (higher values are better for both precision and recall).

To calculate the gradient for predicted judge detention risk in face-space, for any given point in the latent face space (that is, for any given GAN-generated face), we identify the set of GAN-generated images in the neighborhood of the selected point and apply our judge decision predictor (discussed above) to the target face as well as each of the nearby face images. We identify the direction of the gradient in face space, then, as being in the direction of those GAN-generated images that have the largest change in predicted detention likelihood.

C.4 Morphing

As outlined in Subsection A.5, the goal of morphing is to produce two images, x^- and x^+, which have very different predicted probabilities of detention while having very similar visual appearance. Our morphing process uses gradient descent to find these images, and we introduce some variations to this process to produce orthogonalized morphs.

To produce a collection of morph pairs, we first fix a small positive constant α for the step size, and the constants $\breve{m}$ and $\widehat{m}$ required by the definition of the algorithmic hypothesis procedure $\mathcal{P}_m$. We set $\alpha = 0.1$, as this was sufficiently small to ensure that all gradient descent updates decrease the predicted outcome variable when producing x^- (or increased, in for the case of x^+). We set $\breve{m} = 0.1$ and $\widehat{m} = 0.35$, since these values fall in the bottom and top deciles of the predicted values of detention, respectively. To produce a single morph pair (x^-, x^+), we first sample a random seed z_0 from the GAN's latent space. We sampled z_0 following the default approach used by (Karras et al., 2020), including setting the truncation parameter $\psi = 0.5$, as this avoids sampling values for z_0 that are excessively unlikely. To calculate the first image x^-, we let $z^- = z_0$. Given the point z^-, the corresponding synthetic mugshot is $G(z^-)$, and the corresponding predicted detention risk is $m(G(z^-))$. By completing a single forward and backward pass through the composition of both m and G, we can calculate $\nabla m(G(z^-))$, the gradient of predicted detention risk with respect to our current value of z^-. We can then update the value of z^- by subtracting the gradient scaled by the step size:

$$z^- \leftarrow z^- - \alpha \cdot \nabla m(G(z^-)).$$

Since both m and G are differentiable, this reduces the predicted detention risk, provided α is small enough. That is, $m(G(z^-)) < m(G(z_0))$ after a single iteration of the above process. This very similar to the standard gradient descent-based training procedure used for many deep learning models, except that we are updating the input value z^- and keeping the coefficients of m and G fixed. By iterating this process, the value of z^- eventually satisfies $m(G(z^-)) \leq \breve{m}$. Once this condition is satisfied, we terminate the gradient descent process, and set $x^- = G(z^-)$. We employ a similar process to calculate x^+: We set $z^+ = z_0$, reverse the direction of morphing by making the update $\alpha \leftarrow -\alpha$, and iterate the same gradient descent process until $m(G(z^+)) \geq \widehat{m}$. We then set $x^+ = G(z^+)$. The end result is a morphing pair

(x^-, x^+) that satisfies the requirements of $\mathcal{P}_m$.

To produce our orthogonal morphs, we make two variations to the above morphing process. The goal of these variations is to produce a morphing pair (x^-, x^+) that vary by a maximal margin in the outcome dimension (detention risk), while varying by a minimal margin in the x' covariates (well-groomed and skin tone). For the first variation, when running the morphing process, we replace the original model m with a CNN trained on a data set restricted to a sample of observation pairs that match on x' but are discordant in their values of y (which we refer to as our "x'-matched data set"). We also extend the labelling process for skin tone and well-groomed labels by having subjects independently rate the training data set (most of our previous labeling was for images in the validation data set only, since up to this point we did not need labels for training), so that this new CNN can predict both skin tone and well-groomed. We then calculate the values of our morphed points z^- and z^+ in the same manner as above. Since these points are produced with a model that is matched on the x' covariates, $G(z^-)$ and $G(z^+)$ have a smaller difference in predicted covariate values.

However, because of the noise in some of our measures of x', we make an additional variation. For this second variation, given the final values of z^- and z^+, we do a random search in the neighborhood of the new points. We set ε to be one-tenth of the Euclidean distance between z^- and z^+, and sample a series of points z' that are multivariate random normal variables with mean z^+ and standard deviation ε (where each dimension of z' is independent). We continue this sampling until a value of z' is found whose predicted detention risk matches that of z^+ and whose predicted covariate values match those of z^- to a tolerance of 0.001. We then set $x^- = G(z^-)$ and $x^+ = G(z')$. This gives us a morphing pair (x^-, x^+) with a large separation in predicted detention risk, but a small separation in the predicted covariate values. Note that for the first procedure, we use the CNN trained on the x'-matched data set, and for the second procedure we use the original predictive model m. The final result is a pair of mugshots, $G(z^-)$ and $G(z^+)$, one having a high probability of detention, the other a low probability of detention, and each having similar predicted skin tone and similar predicted well-groomed scores. We also address one final subtlety of the specific GAN we use here (styleGAN2). Because this model also infuses some Gaussian noise into various layers of the generator, there are additional free latent variables that can be considered during the morphing process. However, the final stages include a huge number of Gaussian noise variables (up to 512×512 variables). Morphing over all of these variables would allow us to effectively morph the image away from the manifold of images. To solve this, we morph over these noise layers, but with a step size that is reduced by a factor of 100, to avoid large changes. We also use an exponentially decaying step size, to prevent the parameters in these layers from drifting too far from their original values. Finally, we also morph over only the final 7 noise layers, keeping the initial 8 noise layers fixed, since early noise layers can have a larger influence over the appearance of the final face.

C.4.1 A Pseudocode for Morphing

A summary of our morphing algorithm is outlined below in pseudocode format:

Algorithm 1 Targeted face morphing algorithm

Require: StyleGAN2 generator $g : \mathbb{R}^{512} \to \mathbb{R}^{3\times512\times512}$
Require: Detention predictor $m : \mathbb{R}^{3\times512\times512} \to \mathbb{R}$
Require: Covariate predictor $h : \mathbb{R}^{3\times512\times512} \to \mathbb{R}$
Require: Initial input $\boldsymbol{z} \in \mathbb{R}^{512}$
Require: Step size $\alpha \in (0, 1)$
Require: Bound $y^+ \in \mathbb{R}$

Ensure: Final output $\boldsymbol{z} \in \mathbb{R}^{512}$ satisfies $m(g(\boldsymbol{z})) \geq y^+$

function MORPH(g, m, h, $\boldsymbol{z}$, α, y^+)
  repeat
    `// Collect predictions`
    $\boldsymbol{x} \leftarrow g(\boldsymbol{z})$
    $\widehat{y} \leftarrow m(\boldsymbol{x})$
    $\widehat{h} \leftarrow h(\boldsymbol{x})$

    `// Collect gradients`
    $\boldsymbol{\eta}_y = \nabla_{\boldsymbol{z}}\widehat{y}$
    $\boldsymbol{\eta}_h = \nabla_{\boldsymbol{z}}\widehat{h}$
    `// Orthogonalize first argument against the second`
    $\boldsymbol{\eta} = \texttt{Orthogonalize}(\boldsymbol{\eta}_y, \boldsymbol{\eta}_h)$

    `// Update latent vector`
    $\boldsymbol{z} \leftarrow \boldsymbol{z} + \alpha\boldsymbol{\eta}$
  until $\widehat{y} \geq y^+$
  return $\boldsymbol{z}$
end function

D Appendix D: Randomized Lab Experiment

In this appendix we describe the randomized lab experiment we carry out to test the causal relationship between detention decisions and well-groomed and heavy-faced.

The causal interpretation of our new hypotheses is that heavy-faced or well-groomed defendants are released more often because these facial characteristics directly affect how judges form judgments (consciously or unconsciously). Potential confounding arises from the fact that the judge has information that our algorithm does not (as we describe in Section 3), mainly what happens in the hearing itself. Mobius and Rosenblat (2006) show that people's appearance can shape how confident they act, as well as their oral communication skills. Carrying that logic over to our application, it is possible that people who are more heavy-faced or well-groomed either act more confident in court (as signaled by for example their body language, eye contact with the judge or prosecutor, etc.), or are better able to

explain themselves to either the judge or (more likely, since most defendants say little in court at pre-trial detention hearings) their own defense lawyer. These alternate mechanisms are interesting because they suggest different psychologies (and even implicate the psychologies of different people, e.g., the prosecutor or public defender rather than the judge).

We carry out a laboratory experiment that shuts down these two potential channels of confounding to isolate the independent causal effect of defendant appearance on judicial assessments of each defendant's pre-trial risk. At a very high level we carried out two versions of the following experiment, once morphing with respect to well-groomed and once morphing with respect to heavy-faced:

- Describe to subjects the pre-trial system and how the judge must make a decision about who to detain awaiting trial based on a prediction of risk. We then ask them to imagine they are the judge, from different pairs of defendants, which would they be more likely to recommend for detention?
- Subjects are shown 15 defendant pairs as a *training period*. In this stage they are shown actual pairs of mugshots along with structured attributes of each defendant: age, race / ethnicity, the current charge for which the person was arrested, and prior record. After each selection the subject is given feedback about whether the subject chose the defendant at higher risk.
- Subjects are then given 5 minutes to make detention selections without feedback during the *testing period*, and shown information for up to 45 morphed defendant pairs for the well-groomed experiment (randomly selected from a bank of 49 morphed pairs) and similarly up to 45 morphed pairs for the heavy-faced experiment (randomly selected from a bank of 48 morphed pairs). The information shown for each defendant includes the structured variables as described above, as well as synthetic images morphed with respect to either well-groomed or heavy-faced in the direction of higher- or lower risk as described further below. The time limit is intended to mirror the actual decision-making environment of many bond-court environments, where there is not endless amounts of time available to hear each case.

Additional details about the experimental paradigm and analysis include:

- First, we randomly selected 100 synthetic face images from the GAN's latent space
- Second, we randomly assign each synthetic face some values for the structured variables. This is done by extracting real structured-variable values from the actual Mecklenberg dataset (demographics plus current charge plus prior record). We then randomly assign structured variables to synthetic images conditional on the demographics of the structured variables matching the demographics of the synthetic face image. Note this implies that current charge and prior record is not truly random across all face images, but that does not pose a problem given our experimental design.
- We randomly pair up the synthetic defendants. We do this by randomly ordering the synthetic images and their associated structured variables and pairing them up in that order. Let (s) index synthetic pairs. The outcome variable we will analyze below has $y_{is} = 1$ if the study subject (i) chooses to detain the defendant that has the lower of the randomly-assigned order numbers within pair (s); for convenience call that the "top" defendant and the defendant ranked below in the pair the "bottom" defendant.
- For each novel facial feature (well-groomed and heavy-faced), we create two variants of each synthetic image pair (s). One variant morphs the top defendant's image along

the gradient of our feature in the direction towards *lower* risk, and morphs the bottom defendant's image along the gradient of the feature towards *higher* risk, indicated by $v_s = 1$. For the second variant, $v_s = 0$, we do the reverse: morph the top image towards higher risk and the bottom image towards lower risk.

- For each study subject, we randomly select 45 of the 50 defendant pairs to show them (randomly ordered on a per-subject basis), and for each defendant pair, we randomize which variant of the defendant pair they are shown.

We enrolled a total of 500 study subjects on the Prolific platform for the well-groomed experiment, and another 500 subjects for the heavy-faced experiment. We limited participation to US-based study subjects and limited our release for data collection to business hours (US time zones). We offered subjects $2.00 up-front participation incentive plus $0.05 incentive per correct guess during the main evaluation data collection stage. On average subjects in the well-groomed experiment considered 36.5 morphed pairs each, while the figure is 37.1 for the heavy-faced version of the experiment. Our dataset is structured at the level of the respondent-and-defendant-pair, so this leaves us with a total of $18,269$ observations for the well-groomed experiment and $18,548$ observations for the heavy-faced experiment.

Our estimating equation is given as follows, with δ_s a set of defendant-pair fixed effects:

$$y_{is} = \gamma_0 + \gamma_1 v_i s + \delta_s + \epsilon_{is}$$

For our statistical analysis, we cluster the standard errors by respondent (similar results hold if we cluster by respondent and image-pair using the approach from Cameron et al. (2011)). Conditioning on participant fixed effects yields very similar results.

We find that subjects use the structured variables in a way that is consistent with both selecting defendants at higher risk for re-arrest and also consistent with the judge's own use of those variables. The share of subjects who select the defendant within each pair whose structured variables put them at higher risk for re-arrest was 65.6% in the well-groomed version of the experiment and 58.7% in the heavy-faced experiment (as a reminder 50% is the random guessing benchmark). The share of subjects who select the defendant whose structured variables put them at elevated odds of having been detained by the judge equals 70.1% in the well-groomed experiment and 63.1% in the heavy-faced experiment. This tells us not only that the study subjects are taking the task seriously on average (they are not all just guessing randomly), but also that they are making sensible use of the structured variables in this experimental paradigm.

At the same time we also find subjects respond to the random morphings of the defendant faces, above and beyond the effects of the structured variables, as seen in Appendix Table A.XVII. Defendants are 1.3 percentage points more likely to recommend for detention the relatively more well-groomed defendant's image ($p = 0.055$) and 1.9 percentage points more likely for the more heavy-faced image ($p < 0.01$). The table shows that the results are not sensitive to conditioning on study subject fixed effects, which if anything slightly increase the magnitude of our point estimates while shrinking slightly our standard errors (and so together reducing the p-values for our estimates).

It is important to understand what our causal experiment is and is not isolating. Our morphs try to hold other features of these faces constant besides heavy-faced and well-groomed, but visual inspection makes clear that these two novel facial features are also unavoidably correlated to some degree with other aspects of a defendant's face. Given our

data, making such distinctions is difficult; fully teasing these apart might require something like a field experiment inside a local jail that provides grooming assistance to defendants before they walk into court, which is beyond the scope of our analysis here. But from a pragmatic perspective, the exact mechanism may be less relevant given the inequity of the outcome.[91] These mechanisms—aspects of appearance correlated with heavy-facedness or well-groomedness—do sit in a similar orbit with each other. These are "confounders" but they do not suggest radically different explanations for the larger pattern of results.

Other caveats worth keeping in mind include the fact that our study subjects are Prolific workers, not judges. Moreover our subjects are making these decisions in a very different context from which the judges make actual detention decisions. These results should not be considered a substitute for a full-fledged randomized field experiment, but rather might be considered instead another input into the decision a researcher might make about whether to incur the costs of causal testing for our two novel hypotheses.

While these findings are mainly intended to qualitatively establish some relationship, it is perhaps worth noting that the magnitudes implied by our analysis are not trivial. With our randomized morphing procedure, the contrast between the two images the subject sees is on average 3.7 standard deviations different with respect to well-groomed (where the standard deviation in well-groomed is calculated for the validation subsample). For the full-faced version of the experiment, the average image contrast is 4.4 standard deviations. So the subject is essentially selecting which defendant to detain comparing images at the bottom versus the top of the well-groomed (or heavy-faced) distributions. As a benchmark, we can compare the effect of the image to that of the structured variables (current charge, prior record), which as a reminder were randomly assigned to images conditional on race, sex, and age. We statistically relate these structured variables to re-arrest risk among the actual sample of Mecklenburg County defendants, so for each hypothetical defendant in the causal experiment we can calculate the predicted re-arrest risk implied by their structured variables. We calculate that a defendant with structured variables that put them at the top decile of the predicted re-arrest risk distribution is 31 percentage points more likely to be selected for detention by the subjects compared to a defendant in the bottom decile of the predicted re-arrest distribution. So moving along the full distribution of well-groomed or heavy-faced has 4.2% and 6.1% of the effect of moving along the full distribution of re-arrest risk, or equivalently, equal to about a 4 and 6 percentile point movement within the re-arrest risk distribution.[92]

[91] Recall the discussion in Section 4.2 argues against the possibility that these facial characteristics are proxies for risk.

[92] We calculate the effect of re-arrest risk on the subject's detention recommendation through a separate analysis where we assign a $+1$ value if the LHS image is in the top decile of predicted re-arrest risk or the RHS image is in the bottom decile of predicted re-arrest risk, and -1 if the reverse situation is true, 0 else. The effect on subject decisions from moving across the entire predicted risk distribution is twice the coefficient on this variable.

Appendix Tables

Table A.I: Sample construction steps and data missingness filters

Procedure / Data	**Relevant Sample Size**	**Notes**
Raw Data	81166	Total number of arrests downloaded from Mecklenburg County, NC Sheriff's Office public website from January 18, 2017 through January 17, 2020
Filters		
Non-arrest	(8312)	These arrest cases either pertain to probation and parole violations that do not result in new bookings, or can reflect more serious apprehensions pursuant to federal warrants. They do not involve any local pre-trial detention adjudications.
Missing case info	(6238)	Arrests without court case IDs on at least one arrest charge, which means we cannot link arrests to judge pre-trial detention decisions.
Outside observation window	(4737)	The arrest data is matched with inmate data and court record data. These all come from different observation windows. We only consider arrests that fall within the observation window of all three datasets.
Arrested during jail term	(3218)	The arrest date occurs at a time when the individual is already in jail (e.g., due to an offense against another inmate or guard), which typically means pre-trial hearing results in detention – so the judge decision is quite different from out-of-jail arrests.
All cases disposed within 3 days	(2229)	Court cases which are disposed very quickly (within three days). For cases dismissed within such a relatively short time frame, it is likely that judge detention decisions are influenced by a knowledge that dismissal is likely.
Arrested after disposal	(1072)	Arrests with a disposal date occurring earlier than arrest date. This appears to arise from a data recording error.
Charges missing	(542)	These records have no charges listed on the MCSO website in the arrest search. We omit them because we cannot define all outcomes without charges.
Missing inmate dates	(266)	Arrests with a linked inmate record that has missing committed and released date fields. These entries are removed, as we cannot produce all outcomes reliably.
Missing mugshot	(71)	The records with a missing mugshot on MCSO website
Prisoner level separation	(2730)	Since partitioning is implemented at the arrest level, we avoid data spillage at the prisoner level by removing prisoners in the lock-box set who also have an arrest record in the training set or the validation set.
Relevant Sample	51751	
Stratified Sample Partitioning		
Train Set	23138	This is the set on which we trained our judge prediction algorithm.
Validation Set	9604	We use this set to report out-of-sample performance in this draft.
Untouched Lock-Box Set	19009	The untouched data we have set aside for measuring the model's final performance.

Notes: The table above reports how we construct our working data sets by applying various filters during the pre-processing stage.

Table A.II: Test of balance between training dataset and validation dataset

	Train Set	**Validation Set**	**Pairwise comparison p-value**
Sample Size	23138	9604	
Outcome			
Judge detain defendant	0.234	0.233	0.811
Defendant re-arrested before trial	0.250	0.251	0.836
Defendant Characteristics			
Age	31.859	31.631	0.103
Male	0.789	0.782	0.184
White	0.279	0.274	0.443
Black	0.693	0.695	0.783
Other	0.028	0.031	0.205
Arrest Charge			
Violent	0.343	0.343	0.990
Property	0.324	0.317	0.234
Drug	0.204	0.207	0.504
Gun	0.079	0.084	0.106
Other	0.264	0.264	0.981
Arrest Charge Severity			
Felony	0.422	0.428	0.292
Non-Felony	0.578	0.572	
Defendant Prior Record			
Any Prior Conviction	0.463	0.458	0.433
Prior Felony Conviction	0.334	0.328	0.324
Prior Non-Felony Conviction	0.318	0.318	0.979

Notes: This table reports descriptive statistics for our full data set and analysis subsets, which covers the period January 18, 2017, through January 17, 2020, from Mecklenburg County, NC. The untouched holdout data set consists of data from the last 6 months of our study period (July 17, 2019, through January 17, 2019) plus a subset of cases through July 16, 2019, selected by randomly selecting arrestees. The remainder of the data set is then randomly assigned by arrestee to our training data set (used to build our algorithms) or our validation set (on which we report results in this paper draft). Once the paper is accepted, we will report final results for the untouched data set. For additional details of our data filters and partitioning procedures, see Table A.I. We define pre-trial release as being released on the defendant's own recognizance (ROR) or having been assigned and then posting cash bail requirements within three days of arrest. We define re-arrest as experiencing a new arrest before adjudication of the focal arrest, with detained defendants being assigned 0 values for purposes of this table. The pairwise comparison p-value comes from calculating a t-test statistic for the null hypothesis of equivalence of means for a given variable (described by each row label) between the training data set and the validation data set.

P-Values: *p<0.1; **p<0.05; ***p<0.01

Table A.III: Human Intelligence Tasks

Common Name	Survey Number	Short Description	Final Dataset	Subjects	Compensation	Additional Notes
Qualifying task	1	Subjects label 25 images across known variables, in order to identify high-quality raters.	Ratings from several MTurk workers, used to identify a qualified subpopulation of 343 MTurk workers.	600 MTurk Workers	8¢ per image	This survey was periodically re-run when a larger or updated population of raters was required. In total, 343 qualified MTurk Workers were identified across all qualification surveys.
Data collection labelling task part A	2	Subjects label 25 images on sliders for attractiveness, dominance, competence, trustworthiness and well-groomed, a free text input for age, a swatch for skin tone, and a selection for race.	Labels for 32881 images. Includes at least one label for age, race, and skin tone for all images in training and validation, at least three labels for all sliders in training dataset, and at least five labels for all sliders in validation dataset.	343 Qualified MTurk Workers	8¢ per image	The results from all labelling surveys was combined to produce a single image-label dataset.
Data collection labelling task part B	3	Subjects label 25 images on sliders for attractiveness, dominance, competence, trustworthiness, well-groomed, heavy-faced, and potentially other features.	See above.	343 Qualified MTurk Workers	5¢ per image	The results from this survey were merged with the other image label datasets.
Afro-centric features	4	Format is similar to survey 3, but sliders shown are for afrocentric features.	See above.	35 MTurk Workers from qualified population	5¢ per image	Workers were informed that the HIT contained "sensitive material". The results from this survey were merged with the other image label datasets.
Labelling quality check	5	Format is identical to survey 2, but each hit is repeated with multiple subjects.	100 images each with 10 labels.	40 MTurk Workers from qualified population	5¢ per image	The results from this survey were merged with the other image label datasets.
Human guess labelling task	6	Subjects are presented with 50 pairs of mugshots, and instructed to select which person they believe was detained. They are given feedback after each selection (so they can learn to identify patterns), and paid a 5 cent incentive for every correct guess. Each pair is matched to contain the same age bin, race, and sex.	Human guesses for 29,750 image pairs. The final dataset has at least three guesses for 8,001 images, with average of 7.4 guesses per image. Coverage is 79% for images.	595 Prolific Workers	$3.00 base rate, plus a bonus of 5¢ for every correct guess	Because image pairs are matched on age bins, race, and sex, about 21 percent of our validation images do not have a proper match, and hence do not receive a human guess feature.
Morph labelling (along detention gradients)	7	Format and incentive is identical to survey 6, but image pairs shown are all morphed pairs with a high/low detain probability.	Comments described interpreted difference in image pairs, as seen by each Prolific worker. Also, guesses from each prolific worker to get a global masurement of accuracy.	54 Prolific Workers	$3.00 base rate, plus a bonus of 5¢ for every correct guess	
Morph labelling (along residual gradients)	8	Format and incentive is identical to survey 6, but image pairs shown are all morph pairs with a high/low detain probability, and a similar estimated skin tone and well-groomed score.	Comments described interpreted difference in image pairs, as seen by each Prolific worker. Also, guesses from each prolific worker to get a global measurement of accuracy.	52 Prolific Workers	$3.00 base rate, plus a bonus of 5¢ for every correct guess	
Morph labelling (along age gradients)	9	Format and incentive is identical to survey 6, but image pairs shown are all morph pairs with a high/low estimated age. Participants are not told what the "hidden characteristic" is, and must identify it from feedback.	Comments described interpreted difference in image pairs, as seen by each Prolific worker. Also, guesses from each prolific worker to get a global masurement of accuracy.	52 Prolific Workers	$3.00 base rate, plus a bonus of 5¢ for every correct guess	
Data collection labelling task part C	10	Similar to Surveys 2 and 3; subjects label 25 images on slides for mental illness, socioeconomic status, and baby-faced	Labels for 9604 images. Includes at least three labels per image for all images in the validation set.	42 MTurk Workers from qualified population	4.8¢ - 5¢ per image, depending on number of sliders	The results from this survey were merged with the other image label data sets.
Laboratory experiment (well-groomed and heavy-faced)	11	Subjects are presented with pairs of arrest records containing mugshots and information about the defendant's criminal history, charges, age and race. They select which person should be detained based on their risk of re-arrest. After a training phase of 15 pairs with feedback, subjects complete up to 48 selections without feedback within a 5-minute time limit as an evaluation phase. During the evaluation phase, each pair has been morphed so that one randomly selected mugshot exhibits a novel feature (well-groomed or heavy-faced) more strongly, with the other mugshot morphed in the opposite direction.	During the evaluation phase, we collected a total of 18268 and 18548 selections for well-groomed and heavy-faced respectively, based on 96 different pairs of arrest records. The 96 pairs come from 48 different pairs of arrest records, with two variations depending on which mugshot is selected for morphing up versus down.	1000 Prolific Workers (500 per feature)	$2.00 base rate, plus a bonus of 5¢ for every selection that matches a linear regression predicting the riskier defendant.	

Notes: The table above provides a short description of different rounds of data collection via human intelligence tasks. It specifies the objectives and the procedure of each task as well as its incentive structure.

Table A.IV: Summary statistics for human-labeled known facial features from existing psychological research

	Mean Label Value				
Population	Attractiveness	Competence	Dominance	Trustworthiness	Human Guess
Full Sample	3.827	3.792	4.255	3.221	0.496
Race:					
Black	3.831	3.810	4.318	3.245	0.496
White	3.786	3.728	4.106	3.137	0.494
Asian	3.708	3.801	3.819	3.312	0.500
Indian	4.388	4.024	4.012	3.600	0.500
Unknown	4.251	4.031	4.299	3.443	0.505
Age Groups:					
< 25	4.167	3.902	4.193	3.363	0.495
$25 < X < 34$	3.904	3.833	4.284	3.202	0.497
> 34	3.451	3.657	4.284	3.108	0.496
Detained:					
True	3.753	3.704	4.283	3.124	0.511
False	3.850	3.819	4.246	3.250	0.491

Notes: This table shows mean values for each sample sub-group defined at left (row labels) for each human-rated psychological feature indicated in the column heading. Rating ranges were from 1 (low) to 9 (high). Standard deviations of the above labels measured on the full sample size are as follows: attractiveness (0.923), competence (0.911), dominance (0.947), and trustworthiness (0.844). Ratings were conducted on face images (mugshots) taken from Mecklenburg County, NC Sheriff's Office public website. Ratings of attractiveness, competence, dominance and trustworthiness come from subject ratings of mugshot images (see text). Human guess variable comes from showing subjects pairs of mugshot images and asking subjects to identify the defendant they think the judge would be more likely to detain.

Table A.V: Human labeled features for ethnicity and stereotypically Black appearance

	Dependent variable:			
	Algo Judge Detain Prediction		Judge Detain Decision	
	(1)	(2)	(3)	(4)
Male	.1168***	.1149***	.1022***	.0260**
	(.0025)	(.0025)	(.0106)	(.0117)
Age	.0006***	.0003***	−.0008**	−.0014***
	(.0001)	(.0001)	(.0004)	(.0004)
Asian	.0048	.0028	−.0086	−.0146
	(.0045)	(.0045)	(.0193)	(.0191)
Black	.0080**	.0034	−.0013	−.0135
	(.0036)	(.0036)	(.0152)	(.0153)
Hispanic	.0061	.0045	−.0175	−.0241
	(.0043)	(.0043)	(.0184)	(.0182)
Indigenous American	.0089	.0063	.0097	.0003
	(.0095)	(.0094)	(.0403)	(.0398)
Stereotypically Black Appearance	.0004	−.0018**	.0001	−.0037
	(.0006)	(.0008)	(.0027)	(.0034)
Skin-Tone		−.0288***		−.0466*
		(.0062)		(.0262)
Attractiveness		−.0050***		−.0011
		(.0016)		(.0067)
Competence		−.0087***		−.0146**
		(.0017)		(.0072)
Dominance		.0030**		.0058
		(.0012)		(.0051)
Trustworthiness		−.0042**		−.0094
		(.0016)		(.0070)
Human Guess		.0407***		.0851***
		(.0062)		(.0265)
Algo Judge Detain Prediction				.6240***
				(.0434)
Constant	.1347***	.2059***	.1803***	.1761***
	(.0042)	(.0103)	(.0180)	(.0446)
Observations	9,604	9,604	9,604	9,604
Adjusted R^2	.2014	.2222	.0097	.0369

Notes: The table above presents a summary of the results of main paper Tables II and III using an additional feature introduced in the literature that measures the degree to which a person's facial appearance resembles that of a stereotypically Black person which has been found to be closely connected to sentencing decisions (see Eberhardt et al. (2006)). Moreover, the administrative records of MCSO on race are replaced with human labels which capture perceived racial ethnicity of defendants based on their faces. The data on racial ethnicity and stereotypically Black appearance come from subject ratings of mugshot images (see text). Stereotypically Black appearance is coded from 1 (perceived least stereotypically Black) to 9 (perceived most stereotypically Black). For descriptions of other variables, refer to Tables II and III. Regressions follow a linear probability model and also include indicators for unknown racial ethnicity and unknown gender. The base factor levels for gender and ethnicity are female and Caucasian.

P-Values: *p<0.1; **p<0.05; ***p<0.01

Table A.VI: Sensitivity analysis: Non-parametric specifications for skin-tone and known psychological features

	Dependent variable:				
	Algo Judge Detain Prediction		Judge Detain Decision		
	(1)	(2)	(3)	(4)	(5)
Heavy-Faced		−.0180***		−.0220***	−.0118***
		(.0008)		(.0037)	(.0037)
Well-Groomed		−.0134***		−.0109**	−.0033
		(.0011)		(.0051)	(.0051)
Algo Judge Detain Prediction			.6065***		.5699***
			(.0443)		(.0458)
Male	.1133***	.1120***	.0246**	.0912***	.0274**
	(.0025)	(.0024)	(.0118)	(.0108)	(.0119)
Age	.0003***	.0004***	−.0014***	−.0011***	−.0013***
	(.0001)	(.0001)	(.0004)	(.0004)	(.0004)
Black	−.0223***	−.0243***	−.0557***	−.0716***	−.0578***
	(.0040)	(.0039)	(.0174)	(.0175)	(.0174)
Asian	−.0238**	−.0166	−.0639	−.0714	−.0620
	(.0112)	(.0109)	(.0487)	(.0490)	(.0487)
Indigenous American	.0107	.0011	.0645	.0578	.0571
	(.0234)	(.0226)	(.1014)	(.1022)	(.1014)
Human Guess	.0387***	.0275***	.0840***	.0959***	.0803***
	(.0061)	(.0059)	(.0266)	(.0268)	(.0267)
Constant	.0958***	.2731***	.0118	.2536***	.0980**
	(.0076)	(.0108)	(.0333)	(.0487)	(.0499)
Indicators for Skin-Tone?	YES	YES	YES	YES	YES
Indicators for Psychological Features?	YES	YES	YES	YES	YES
Observations	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	.2496	.2987	.0371	.0224	.0379

Notes: The above table replicates the richest specifications of main paper Tables II, III, V and VI, but now relaxing the linearity assumption for skin tone and known psychological features. The table shows results of estimating a linear probability specification regressing algorithmic prediction of judge detain decision (columns (1) and (2)) and actual judges' detain decision (columns (3) through (5)) against different explanatory variables, using data from the validation set separately for male and female defendants. The Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to validation set observations. Measures of defendant demographics and current arrest charge come from Mecklenburg County administrative data. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance and trustworthiness come from subject ratings of mugshot images (see text). Human guess variable comes from showing subjects pairs of mugshot images and asking subjects to identify the defendant they think the judge would be more likely to detain. The base factor levels for gender and race are female and white. Regression specifications also include indicators for unknown race and unknown gender.

P-Values: *p<0.1; **p<0.05; ***p<0.01

Table A.VII: Cross gender sensitivity analysis: Non-parametric specification for skin-tone and known psychological features

	Dependent variable:									
	Algo Judge Detain Prediction				Judge Detain Decision					
	Male Defendants		Female Defendants		Male Defendants			Female Defendants		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Faced		−.0193***		−.0106***		−.0191***	−.0078*		−.0277***	−.0232***
		(.0010)		(.0014)		(.0043)	(.0044)		(.0066)	(.0067)
Well-Groomed		−.0128***		−.0174***		−.0072	.0002		−.0254***	−.0179*
		(.0013)		(.0020)		(.0060)	(.0059)		(.0097)	(.0098)
Algo Judge Detain Prediction					.6027***		.5814***	.5376***		.4297***
					(.0505)		(.0523)	(.1020)		(.1054)
Age	.0004***	.0006***	−.0003*	−.0004**	−.0014***	−.0010**	−.0014***	−.0012	−.0016*	−.0014*
	(.0001)	(.0001)	(.0002)	(.0002)	(.0005)	(.0005)	(.0005)	(.0008)	(.0008)	(.0008)
Black	−.0028	−.0068	−.0786***	−.0761***	−.0441**	−.0494**	−.0455**	−.1018***	−.1394***	−.1067***
	(.0048)	(.0046)	(.0065)	(.0062)	(.0209)	(.0211)	(.0209)	(.0309)	(.0298)	(.0308)
Asian	−.0091	−.0025	−.0625***	−.0544***	−.0536	−.0543	−.0528	−.0915	−.1106	−.0872
	(.0129)	(.0124)	(.0209)	(.0202)	(.0560)	(.0565)	(.0561)	(.0963)	(.0962)	(.0960)
Indigenous American	.0173	.0087	−.0169	−.0251	−.0782	−.0780	−.0831	.3193**	.2876**	.2984**
	(.0300)	(.0290)	(.0316)	(.0306)	(.1307)	(.1318)	(.1307)	(.1456)	(.1458)	(.1452)
Human Guess	.0348***	.0247***	.0438***	.0281**	.0678**	.0809***	.0665**	.1573***	.1512***	.1391**
	(.0069)	(.0067)	(.0120)	(.0117)	(.0303)	(.0306)	(.0303)	(.0556)	(.0558)	(.0556)
Constant	.1849***	.3630***	.1516***	.3190***	.0484	.3024***	.0914	−.0064	.3863***	.2492***
	(.0089)	(.0126)	(.0133)	(.0189)	(.0399)	(.0572)	(.0598)	(.0630)	(.0902)	(.0959)
Indicators for Skin-Tone?	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Indicators for Psychological Features?	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	7,511	7,511	2,092	2,092	7,511	7,511	7,511	2,092	2,092	2,092
Adjusted R^2	.0783	.1395	.1990	.2542	.0264	.0106	.0266	.0482	.0477	.0550

Notes: The above table replicates the richest specifications of main paper Tables II, III, V, and VI, but now relaxing the linearity assumption for skin tone and psychological features while introducing low-level interactions with defendant's gender. The table shows results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision (columns (1) through (4)) and actual judges' detain decision (columns (5) through (10)) against different explanatory variables, using data from the validation set separately for male and female defendants. Algorithmic predictions of judges' decisions come from applying algorithm built with face images in the training data set to validation set observations. Data on well-groomed, skin tone, and psychological features (i.e., attractiveness, competence, dominance, and trustworthiness) come from subject ratings of mugshot images (see text). Human guess variable comes from showing subjects pairs of mugshot images and asking subjects to identify the defendant they think the judge would be more likely to detain. Regression specifications also include indicators for unknown race.

P-Values: *p<0.1; **p<0.05; ***p<0.01

Table A.VIII: Cross race sensitivity analysis: Non-parametric specification for skin-tone and known psychological features

	Dependent variable:									
	Algo Judge Detain Prediction				Judge Detain Decision					
	Black Defendants		Non-Black Defendants		Black Defendants			Non-Black Defendants		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Faced		−.0174***		−.0166***		−.0210***	−.0112**		−.0214***	−.0126*
		(.0010)		(.0014)		(.0044)	(.0045)		(.0067)	(.0068)
Well-Groomed		−.0184***		−.0046**		−.0111*	−.0008		−.0114	−.0090
		(.0014)		(.0019)		(.0062)	(.0063)		(.0090)	(.0090)
Algo Judge Detain Prediction					.5915***		.5602***	.5690***		.5270***
					(.0532)		(.0553)	(.0852)		(.0874)
Male	.1442***	.1415***	.0592***	.0607***	.0435***	.1245***	.0453***	−.0086	.0276	−.0045
	(.0031)	(.0030)	(.0040)	(.0039)	(.0154)	(.0135)	(.0155)	(.0189)	(.0183)	(.0190)
Age	.0005***	.0005***	−.0002	−.00003	−.0013***	−.0010**	−.0013***	−.0015**	−.0015*	−.0015*
	(.0001)	(.0001)	(.0002)	(.0002)	(.0005)	(.0005)	(.0005)	(.0008)	(.0008)	(.0008)
Human Guess	.0328***	.0224***	.0467***	.0349***	.0737**	.0846***	.0720**	.1037**	.1124**	.0940*
	(.0072)	(.0070)	(.0111)	(.0109)	(.0312)	(.0315)	(.0313)	(.0510)	(.0515)	(.0513)
Constant	.0514***	.2545***	.1445***	.2632***	.2020***	.4109***	.2683***	.0250	.2961***	.1574*
	(.0172)	(.0191)	(.0121)	(.0182)	(.0745)	(.0865)	(.0870)	(.0567)	(.0858)	(.0884)
Indicators for Skin-Tone?	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Indicators for Psychological Features?	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	6,673	6,673	2,931	2,931	6,673	6,673	6,673	2,931	2,931	2,931
Adjusted R^2	.3146	.3649	.1423	.1850	.0407	.0266	.0413	.0303	.0194	.0313

Notes: The above table replicates the richest specifications of main paper Tables II, III, V, and VI, but now relaxing the linearity assumption for skin tone and psychological features while introducing low-level interactions with defendant's race. The table shows results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision (columns (1) through (4)) and actual judges' detain decision (columns (5) through (10)) against different explanatory variables, using data from the validation set separately for Black and non-Black defendants. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to validation set observations. Data on well-groomed, skin tone, and psychological features (i.e., attractiveness, competence, dominance and trustworthiness) come from subject ratings of mugshot images (see text). Human guess variable comes from showing subjects pairs of mugshot images and asking subjects to identify the defendant they think the judge would be more likely to detain. Regression specifications also include indicators for unknown race.

P-Values: *p<0.1; **p<0.05; ***p<0.01

Table A.IX: Relationship between novel features and algorithm's prediction controlling for indicators of defendant drug involvement

	Dependent variable:						
	Algo Judge Detain Prediction						
						Drug Possession Charge	No Drug Possession Charge
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	−0.0181***	−0.0189***			−0.0182***	−0.0167***	−0.0184***
	(0.0009)	(0.0008)			(0.0008)	(0.0021)	(0.0009)
Well-Groomed			−0.0172***	−0.0153***	−0.0133***	−0.0098***	−0.0141***
			(0.0011)	(0.0012)	(0.0012)	(0.0028)	(0.0013)
Male		0.1117***		0.1155***	0.1130***	0.0980***	0.1151***
		(0.0024)		(0.0025)	(0.0024)	(0.0069)	(0.0026)
Age		0.0004***		0.0002**	0.0004***	0.0002	0.0004***
		(0.0001)		(0.0001)	(0.0001)	(0.0003)	(0.0001)
Black		−0.0187***		−0.0168***	−0.0183***	−0.0119	−0.0194***
		(0.0035)		(0.0036)	(0.0035)	(0.0087)	(0.0038)
Asian		−0.0187*		−0.0160	−0.0140	0.0088	−0.0184
		(0.0111)		(0.0113)	(0.0110)	(0.0292)	(0.0119)
Indigenous American		−0.0006		0.0172	0.0040	0.0167	0.0002
		(0.0232)		(0.0236)	(0.0230)	(0.0527)	(0.0255)
Skin-Tone		−0.0453***		−0.0440***	−0.0472***	−0.0387***	−0.0489***
		(0.0057)		(0.0058)	(0.0056)	(0.0139)	(0.0062)
Attractiveness		−0.0086***		0.0008	−0.0033**	−0.0068*	−0.0028
		(0.0015)		(0.0016)	(0.0016)	(0.0038)	(0.0017)
Competence		−0.0085***		−0.0060***	−0.0061***	−0.0093**	−0.0055***
		(0.0016)		(0.0017)	(0.0016)	(0.0040)	(0.0018)
Dominance		0.0059***		0.0031***	0.0058***	0.0064**	0.0057***
		(0.0012)		(0.0012)	(0.0012)	(0.0028)	(0.0013)
Trustworthiness		−0.0014		−0.0024	0.00001	0.0018	−0.0002
		(0.0016)		(0.0016)	(0.0016)	(0.0040)	(0.0017)
Human Guess		0.0336***		0.0339***	0.0286***	0.0170	0.0308***
		(0.0061)		(0.0062)	(0.0060)	(0.0143)	(0.0067)
Drug Possession	0.0049	−0.0020	0.0073**	−0.0006	−0.0027		
	(0.0031)	(0.0027)	(0.0031)	(0.0028)	(0.0027)		
Constant	0.3474***	0.3122***	0.3335***	0.2570***	0.3430***	0.3480***	0.3429***
	(0.0051)	(0.0102)	(0.0054)	(0.0099)	(0.0104)	(0.0262)	(0.0114)
Observations	9,604	9,604	9,604	9,604	9,604	1,442	8,162
Adjusted R^2	0.0385	0.2627	0.0251	0.2360	0.2727	0.2014	0.2828

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features to the algorithm's overall prediction of judge detention decisions, with some control for an indicator of the defendant's drug involvement. Specifically we control for whether the defendant's current charge is for drug possession in columns (1) through (5), which use the full validation (test set) sample. In column (7) we re-run the analysis using just those defendants who have some indication of drug involvement, while column (8) uses the remaining sample of defendants.
P-Values: *p<.1; **p<.05; ***p<.01

Table A.X: Relationship between novel features and algorithm's prediction controlling for indicator of defendant's mental health

	Dependent variable:						
	Algo Judge Detain Prediction						
						MI ≥ Median(MI)	MI < Median(MI)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	−0.0175***	−0.0183***			−0.0177***	−0.0190***	−0.0162***
	(0.0009)	(0.0008)			(0.0008)	(0.0011)	(0.0012)
Well-Groomed			−0.0157***	−0.0141***	−0.0126***	−0.0143***	−0.0109***
			(0.0011)	(0.0012)	(0.0012)	(0.0016)	(0.0017)
Male		0.1129***		0.1168***	0.1139***	0.1132***	0.1142***
		(0.0024)		(0.0025)	(0.0024)	(0.0033)	(0.0036)
Age		0.0004***		0.0002**	0.0004***	0.0006***	−0.00003
		(0.0001)		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		−0.0179***		−0.0160***	−0.0178***	−0.0172***	−0.0189***
		(0.0035)		(0.0036)	(0.0035)	(0.0049)	(0.0050)
Asian		−0.0174		−0.0148	−0.0132	−0.0285	−0.0048
		(0.0111)		(0.0113)	(0.0110)	(0.0174)	(0.0141)
Indigenous American		0.0004		0.0175	0.0045	−0.0312	0.0318
		(0.0231)		(0.0235)	(0.0230)	(0.0362)	(0.0296)
Skin-Tone		−0.0443***		−0.0428***	−0.0463***	−0.0468***	−0.0462***
		(0.0057)		(0.0058)	(0.0056)	(0.0079)	(0.0081)
Attractiveness		−0.0076***		0.0013	−0.0029*	−0.0012	−0.0055**
		(0.0015)		(0.0016)	(0.0016)	(0.0022)	(0.0022)
Competence		−0.0077***		−0.0053***	−0.0056***	−0.0072***	−0.0040*
		(0.0016)		(0.0017)	(0.0016)	(0.0023)	(0.0024)
Dominance		0.0053***		0.0025**	0.0054***	0.0063***	0.0050***
		(0.0012)		(0.0012)	(0.0012)	(0.0016)	(0.0017)
Trustworthiness		−0.0011		−0.0021	0.0002	0.0001	0.0001
		(0.0016)		(0.0016)	(0.0016)	(0.0023)	(0.0022)
Human Guess		0.0311***		0.0313***	0.0270***	0.0210**	0.0346***
		(0.0061)		(0.0062)	(0.0060)	(0.0085)	(0.0086)
Mental Illness (MI)	0.0061***	0.0048***	0.0044***	0.0056***	0.0037***		
	(0.0009)	(0.0008)	(0.0009)	(0.0008)	(0.0008)		
Constant	0.3207***	0.2850***	0.3099***	0.2262***	0.3201***	0.3425***	0.3279***
	(0.0064)	(0.0110)	(0.0074)	(0.0108)	(0.0114)	(0.0144)	(0.0154)
Observations	9,604	9,604	9,604	9,604	9,604	5,068	4,536
Adjusted R^2	0.0433	0.2656	0.0270	0.2399	0.2743	0.2746	0.2644

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features to the algorithm's overall prediction of judge detention decisions, with some control for an indicator of the defendant's mental health. Specifically we have a separate sample of study subjects independently rate mugshots in the validation (test set) sample for their perceptions of the mental health of the person, and then control for that in the regressions shown in columns (1) through (5), which use the full validation (test set) sample. In column (6) we re-run the analysis using just those defendants who are above median in their mental illness ratings, while column (7) uses the remaining sample of defendants.
P-Values: *p<.1; **p<.05; ***p<.01

Table A.XI: Relationship between novel features and algorithm's prediction controlling for defendant's perceived socio-economic status (SES)

	Dependent variable:						
	Algo Judge Detain Prediction						
						SES ≥ Median(SES)	SES < Median(SES)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Face	−0.0172***	−0.0180***			−0.0175***	−0.0168***	−0.0186***
	(0.0009)	(0.0008)			(0.0008)	(0.0011)	(0.0013)
Well-Groomed			−0.0135***	−0.0131***	−0.0116***	−0.0097***	−0.0159***
			(0.0011)	(0.0012)	(0.0012)	(0.0015)	(0.0018)
Male		0.1121***		0.1157***	0.1132***	0.1067***	0.1237***
		(0.0024)		(0.0025)	(0.0024)	(0.0031)	(0.0039)
Age		0.0004***		0.0002**	0.0003***	0.0002	0.0005***
		(0.0001)		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		−0.0228***		−0.0211***	−0.0218***	−0.0198***	−0.0222***
		(0.0035)		(0.0036)	(0.0035)	(0.0043)	(0.0062)
Asian		−0.0195*		−0.0175	−0.0153	−0.0074	−0.0359*
		(0.0110)		(0.0112)	(0.0110)	(0.0130)	(0.0204)
Indigenous American		0.0001		0.0166	0.0039	0.0115	−0.0269
		(0.0230)		(0.0234)	(0.0229)	(0.0258)	(0.0482)
Skin-Tone		−0.0397***		−0.0381***	−0.0422***	−0.0438***	−0.0434***
		(0.0057)		(0.0058)	(0.0057)	(0.0070)	(0.0095)
Attractiveness		−0.0063***		0.0021	−0.0021	−0.0035*	−0.0023
		(0.0015)		(0.0016)	(0.0016)	(0.0020)	(0.0026)
Competence		−0.0076***		−0.0055***	−0.0056***	−0.0040*	−0.0081***
		(0.0016)		(0.0017)	(0.0016)	(0.0021)	(0.0026)
Dominance		0.0054***		0.0027**	0.0054***	0.0048***	0.0068***
		(0.0012)		(0.0012)	(0.0012)	(0.0015)	(0.0018)
Trustworthiness		−0.0014		−0.0026	−0.0002	−0.0020	0.0023
		(0.0016)		(0.0016)	(0.0016)	(0.0020)	(0.0026)
Human Guess		0.0299***		0.0307***	0.0262***	0.0309***	0.0207**
		(0.0060)		(0.0062)	(0.0060)	(0.0078)	(0.0095)
Socioeconomic Status (SES)	−0.0146***	−0.0098***	−0.0128***	−0.0100***	−0.0083***		
	(0.0010)	(0.0009)	(0.0010)	(0.0009)	(0.0009)		
Constant	0.4087***	0.3448***	0.3744***	0.2896***	0.3662***	0.3239***	0.3492***
	(0.0064)	(0.0105)	(0.0062)	(0.0103)	(0.0107)	(0.0132)	(0.0171)
Observations	9,604	9,604	9,604	9,604	9,604	5,651	3,953
Adjusted R^2	0.0608	0.2714	0.0408	0.2449	0.2786	0.2504	0.2847

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features to the algorithm's overall prediction of judge detention decisions, with some control for the defendant's socio-economic status (SES). Specifically we have a separate sample of study subjects independently rate mugshots in the validation (test set) sample for their perceptions of the defendant's SES, then control for that in the regressions shown in columns (1) through (5), which use the full validation (test set) sample. In columns (6) we re-run the analysis using just those defendants who are above median in their rated SES, while column (7) uses the remaining sample of defendants.

P-Values: *p<.1; **p<.05; ***p<.01

Table A.XII: Relationship between novel features and algorithm's prediction controlling for defendant's baby-faced feature

	Dependent variable:						
	Algo Judge Detain Prediction						
						BF ≥ Median(BF)	BF < Median(BF)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Face	−0.0156***	−0.0177***			−0.0172***	−0.0136***	−0.0212***
	(0.0009)	(0.0009)			(0.0009)	(0.0011)	(0.0013)
Well-Groomed			−0.0140***	−0.0140***	−0.0128***	−0.0121***	−0.0137***
			(0.0011)	(0.0012)	(0.0012)	(0.0015)	(0.0018)
Male		0.1103***		0.1128***	0.1118***	0.1165***	0.1053***
		(0.0024)		(0.0025)	(0.0024)	(0.0030)	(0.0041)
Age		0.0003***		−0.0001	0.0002**	−0.0004***	0.0008***
		(0.0001)		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		−0.0176***		−0.0151***	−0.0175***	−0.0237***	−0.0094*
		(0.0035)		(0.0036)	(0.0035)	(0.0045)	(0.0056)
Asian		−0.0178		−0.0145	−0.0134	−0.0159	−0.0093
		(0.0111)		(0.0112)	(0.0110)	(0.0139)	(0.0175)
Indigenous American		0.0007		0.0176	0.0048	0.0287	−0.0284
		(0.0231)		(0.0234)	(0.0230)	(0.0271)	(0.0407)
Skin-Tone		−0.0455***		−0.0446***	−0.0473***	−0.0462***	−0.0463***
		(0.0057)		(0.0058)	(0.0056)	(0.0071)	(0.0091)
Attractiveness		−0.0082***		0.0005	−0.0033**	−0.0036*	−0.0019
		(0.0015)		(0.0016)	(0.0016)	(0.0020)	(0.0025)
Competence		−0.0084***		−0.0062***	−0.0061***	−0.0046**	−0.0068***
		(0.0016)		(0.0017)	(0.0016)	(0.0021)	(0.0026)
Dominance		0.0054***		0.0025**	0.0054***	0.0062***	0.0052***
		(0.0012)		(0.0012)	(0.0012)	(0.0015)	(0.0018)
Trustworthiness		−0.0009		−0.0015	0.0003	−0.0001	−0.0010
		(0.0016)		(0.0016)	(0.0016)	(0.0020)	(0.0025)
Human Guess		0.0327***		0.0322***	0.0281***	0.0241***	0.0317***
		(0.0061)		(0.0062)	(0.0060)	(0.0077)	(0.0095)
Baby-Faced (BF)	−0.0133***	−0.0052***	−0.0141***	−0.0092***	−0.0042***		
	(0.0010)	(0.0010)	(0.0010)	(0.0010)	(0.0010)		
Constant	0.3897***	0.3325***	0.3770***	0.3006***	0.3578***	0.3264***	0.3510***
	(0.0058)	(0.0108)	(0.0061)	(0.0109)	(0.0110)	(0.0136)	(0.0161)
Observations	9,604	9,604	9,604	9,604	9,604	5,250	4,354
Adjusted R^2	0.0563	0.2650	0.0446	0.2433	0.2741	0.2957	0.2256

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features to the algorithm's overall prediction of judge detention decisions, with some control for the defendant's degree of baby-facedness. Specifically we have a separate sample of study subjects independently rate mugshots in the validation (test set) sample based on their relative baby-faced looks, then control for that in the regressions shown in columns (1) through (5), which use the full validation (test set) sample. In columns (6) we re-run the analysis using just those defendants who are above median in their baby-faced ratings, while column (7) uses the remaining sample of defendants.

P-Values: *p<.1; **p<.05; ***p<.01

Table A.XIII: Relationship between novel features and judge decision controlling for indicators of defendant drug involvement

	Dependent variable:									
	Judge Detain Decision									
							Drug Possession Charge		No Drug Possession Charge	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Faced	−0.0237***	−0.0227***			−0.0221***		−0.0179*		−0.0225***	
	(0.0036)	(0.0036)			(0.0037)		(0.0094)		(0.0040)	
Well-Groomed			−0.0199***	−0.0128**	−0.0103**		−0.0117		−0.0100*	
			(0.0043)	(0.0051)	(0.0051)		(0.0128)		(0.0056)	
Algo Judge Detain Prediction						0.6172***		0.3612***		0.6552***
						(0.0434)		(0.1163)		(0.0467)
Male		0.0935***		0.0975***	0.0945***	0.0259**	−0.0038	−0.0396	0.1090***	0.0350***
		(0.0108)		(0.0108)	(0.0108)	(0.0117)	(0.0312)	(0.0331)	(0.0115)	(0.0126)
Age		−0.0012***		−0.0014***	−0.0013***	−0.0016***	0.0014	0.0013	−0.0016***	−0.0019***
		(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0012)	(0.0012)	(0.0004)	(0.0004)
Black		−0.0646***		−0.0624***	−0.0643***	−0.0521***	−0.1003**	−0.0966**	−0.0547***	−0.0411**
		(0.0155)		(0.0156)	(0.0155)	(0.0154)	(0.0392)	(0.0391)	(0.0169)	(0.0168)
Asian		−0.0742		−0.0730	−0.0705	−0.0643	−0.2187*	−0.2381*	−0.0503	−0.0390
		(0.0487)		(0.0489)	(0.0488)	(0.0483)	(0.1321)	(0.1312)	(0.0525)	(0.0520)
Indigenous American		0.0495		0.0691	0.0530	0.0575	0.0833	0.0723	0.0468	0.0554
		(0.1019)		(0.1020)	(0.1019)	(0.1010)	(0.2380)	(0.2374)	(0.1125)	(0.1114)
Skin-Tone		−0.1059***		−0.1036***	−0.1074***	−0.0759***	−0.1075*	−0.0911	−0.1054***	−0.0712***
		(0.0250)		(0.0251)	(0.0250)	(0.0249)	(0.0628)	(0.0628)	(0.0273)	(0.0270)
Attractiveness		−0.0082		0.0009	−0.0041	−0.0009	0.0097	0.0109	−0.0072	−0.0037
		(0.0067)		(0.0070)	(0.0070)	(0.0067)	(0.0173)	(0.0165)	(0.0077)	(0.0073)
Competence		−0.0199***		−0.0180**	−0.0181**	−0.0148**	−0.0403**	−0.0382**	−0.0135*	−0.0101
		(0.0072)		(0.0073)	(0.0073)	(0.0072)	(0.0183)	(0.0182)	(0.0079)	(0.0078)
Dominance		0.0113**		0.0079	0.0113**	0.0060	0.0120	0.0076	0.0108*	0.0055
		(0.0052)		(0.0051)	(0.0052)	(0.0051)	(0.0129)	(0.0127)	(0.0056)	(0.0056)
Trustworthiness		−0.0088		−0.0106	−0.0077	−0.0095	−0.0193	−0.0224	−0.0053	−0.0068
		(0.0071)		(0.0071)	(0.0071)	(0.0070)	(0.0183)	(0.0181)	(0.0077)	(0.0076)
Human Guess		0.1032***		0.1057***	0.0993***	0.0861***	0.0723	0.0746	0.1051***	0.0886***
		(0.0267)		(0.0268)	(0.0268)	(0.0265)	(0.0648)	(0.0643)	(0.0294)	(0.0290)
Drug Possession	−0.0206*	−0.0330***	−0.0174	−0.0310**	−0.0336***	−0.0304**				
	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0121)	(0.0119)				
Constant	0.3616***	0.4521***	0.3310***	0.3713***	0.4759***	0.2042***	0.5334***	0.3313***	0.4538***	0.1743***
	(0.0198)	(0.0447)	(0.0210)	(0.0430)	(0.0463)	(0.0416)	(0.1186)	(0.1088)	(0.0502)	(0.0449)
Naive-AUC	0.546	0.605	0.533	0.596	0.605	0.637	0.615	0.624	0.609	0.645
Observations	9,604	9,604	9,604	9,604	9,604	9,604	1,442	1,442	8,162	8,162
Adjusted R^2	0.0044	0.0222	0.0022	0.0189	0.0225	0.0385	0.0210	0.0251	0.0252	0.0439

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features, or the algorithm's overall prediction of judge detention decisions, to actual judge detention decisions, with some control for an indicator of the defendant's drug involvement. Specifically we control for whether the defendant's current charge is for drug possession in columns (1) through (6), which use the full validation (test set) sample. In columns (7) and (8) we re-run the analysis using just those defendants who have some indication of drug involvement, while columns (9) and (10) use the remaining sample of defendants.

P-Values: *p<.1; **p<.05; ***p<.01

Table A.XIV: Relationship between novel features and judge decision controlling for indicator of defendant's mental health

	Dependent variable:									
	Judge Detain Decision						MI ≥ Median(MI)		MI < Median(MI)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Faced	−0.0223***	−0.0214***			−0.0210***		−0.0229***		−0.0190***	
	(0.0036)	(0.0037)			(0.0037)		(0.0051)		(0.0054)	
Well-Groomed			−0.0168***	−0.0106**	−0.0087*		−0.0135*		−0.0039	
			(0.0044)	(0.0052)	(0.0052)		(0.0072)		(0.0075)	
Algo Judge Detain Prediction						0.6109***		0.4845***		0.7695***
						(0.0436)		(0.0602)		(0.0632)
Male		0.0939***		0.0981***	0.0946***	0.0266**	0.0897***	0.0352**	0.1010***	0.0145
		(0.0108)		(0.0108)	(0.0108)	(0.0118)	(0.0147)	(0.0161)	(0.0159)	(0.0173)
Age		−0.0012***		−0.0014***	−0.0012***	−0.0015***	−0.0011*	−0.0014***	−0.0014**	−0.0014**
		(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0006)	(0.0005)	(0.0006)	(0.0006)
Black		−0.0634***		−0.0611***	−0.0633***	−0.0514***	−0.0484**	−0.0409*	−0.0793***	−0.0640***
		(0.0156)		(0.0156)	(0.0156)	(0.0154)	(0.0219)	(0.0218)	(0.0222)	(0.0218)
Asian		−0.0718		−0.0707	−0.0689	−0.0623	−0.0484	−0.0385	−0.0850	−0.0807
		(0.0488)		(0.0489)	(0.0488)	(0.0484)	(0.0775)	(0.0772)	(0.0625)	(0.0615)
Indigenous American		0.0505		0.0687	0.0533	0.0575	0.0472	0.0596	0.0551	0.0387
		(0.1019)		(0.1020)	(0.1019)	(0.1010)	(0.1614)	(0.1607)	(0.1308)	(0.1287)
Skin-Tone		−0.1047***		−0.1019***	−0.1061***	−0.0754***	−0.0810**	−0.0554	−0.1354***	−0.0994***
		(0.0250)		(0.0251)	(0.0250)	(0.0249)	(0.0352)	(0.0351)	(0.0357)	(0.0352)
Attractiveness		−0.0070		0.0012	−0.0037	−0.0002	−0.0048	−0.0045	−0.0029	0.0043
		(0.0068)		(0.0070)	(0.0070)	(0.0067)	(0.0100)	(0.0095)	(0.0099)	(0.0093)
Competence		−0.0183**		−0.0165**	−0.0169**	−0.0136*	−0.0222**	−0.0201**	−0.0110	−0.0072
		(0.0072)		(0.0073)	(0.0073)	(0.0072)	(0.0102)	(0.0101)	(0.0104)	(0.0102)
Dominance		0.0101*		0.0067	0.0101*	0.0052	0.0178**	0.0123*	0.0016	−0.0032
		(0.0052)		(0.0052)	(0.0052)	(0.0051)	(0.0072)	(0.0071)	(0.0075)	(0.0074)
Trustworthiness		−0.0081		−0.0099	−0.0072	−0.0088	−0.0058	−0.0086	−0.0099	−0.0103
		(0.0071)		(0.0071)	(0.0071)	(0.0070)	(0.0102)	(0.0101)	(0.0099)	(0.0097)
Human Guess		0.0986***		0.1009***	0.0958***	0.0824***	0.0991***	0.0957**	0.0929**	0.0672*
		(0.0268)		(0.0268)	(0.0268)	(0.0266)	(0.0380)	(0.0378)	(0.0378)	(0.0373)
Mental Illness (MI)	0.0103***	0.0073**	0.0088**	0.0088**	0.0065*	0.0055				
	(0.0033)	(0.0035)	(0.0034)	(0.0035)	(0.0035)	(0.0034)				
Constant	0.3099***	0.4032***	0.2783***	0.3165***	0.4276***	0.1724***	0.4530***	0.2076***	0.4560***	0.1821***
	(0.0248)	(0.0486)	(0.0285)	(0.0469)	(0.0507)	(0.0445)	(0.0643)	(0.0593)	(0.0680)	(0.0582)
Naive-AUC	0.548	0.602	0.535	0.594	0.602	0.636	0.598	0.613	0.605	0.663
Observations	9,604	9,604	9,604	9,604	9,604	9,604	5,068	5,068	4,536	4,536
Adjusted R^2	0.0051	0.0219	0.0027	0.0188	0.0220	0.0381	0.0200	0.0277	0.0212	0.0498

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features, or the algorithm's overall prediction of judge detention decisions, to actual judge detention decisions, with some control for an indicator of the defendant's mental health. Specifically we have a separate sample of study subjects independently rate mugshots in the validation (test set) sample for their perceptions of the mental health of the person, and then control for that in the regressions shown in columns (1) through (6), which use the full validation (test set) sample. In columns (7) and (8) we re-run the analysis using just those defendants who are above median in their mental illness ratings, while columns (9) and (10) use the remaining sample of defendants.

P-Values: *p<.1; **p<.05; ***p<.01

Table A.XV: Relationship between novel features and judge decision controlling for defendant's perceived socioeconomic status (SES)

	Dependent variable:									
	Judge Detain Decision						SES ≥ Median(SES)		SES < Median(SES)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Face	−0.0220***	−0.0208***			−0.0205***		−0.0163***		−0.0267***	
	(0.0036)	(0.0037)			(0.0037)		(0.0047)		(0.0058)	
Well-Groomed			−0.0143***	−0.0086*	−0.0067		−0.0053		−0.0114	
			(0.0044)	(0.0052)	(0.0052)		(0.0066)		(0.0083)	
Algo Judge Detain Prediction						0.6012***		0.6809***		0.5040***
						(0.0438)		(0.0564)		(0.0690)
Male		0.0927***		0.0963***	0.0934***	0.0267**	0.0890***	0.0168	0.1001***	0.0408**
		(0.0107)		(0.0107)	(0.0107)	(0.0118)	(0.0135)	(0.0147)	(0.0177)	(0.0196)
Age		−0.0012***		−0.0014***	−0.0012***	−0.0015***	−0.0016***	−0.0018***	−0.0009	−0.0011*
		(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0005)	(0.0005)	(0.0006)	(0.0006)
Black		−0.0713***		−0.0698***	−0.0707***	−0.0572***	−0.0504***	−0.0368**	−0.1046***	−0.0915***
		(0.0156)		(0.0157)	(0.0156)	(0.0155)	(0.0187)	(0.0185)	(0.0278)	(0.0278)
Asian		−0.0750		−0.0751	−0.0725	−0.0649	−0.1278**	−0.1233**	0.0499	0.0663
		(0.0487)		(0.0488)	(0.0488)	(0.0483)	(0.0570)	(0.0562)	(0.0919)	(0.0915)
Indigenous America		0.0501		0.0673	0.0524	0.0570	0.1625	0.1587	−0.3077	−0.2893
		(0.1018)		(0.1020)	(0.1018)	(0.1010)	(0.1136)	(0.1122)	(0.2171)	(0.2163)
Skin-Tone		−0.0969***		−0.0936***	−0.0984***	−0.0705***	−0.0794***	−0.0492	−0.1419***	−0.1119***
		(0.0251)		(0.0252)	(0.0251)	(0.0249)	(0.0308)	(0.0305)	(0.0430)	(0.0427)
Attractiveness		−0.0046		0.0027	−0.0022	0.0012	−0.0098	−0.0059	0.0052	0.0083
		(0.0068)		(0.0070)	(0.0071)	(0.0067)	(0.0088)	(0.0083)	(0.0118)	(0.0112)
Competence		−0.0180**		−0.0167**	−0.0168**	−0.0135*	−0.0059	−0.0030	−0.0322***	−0.0286**
		(0.0072)		(0.0073)	(0.0073)	(0.0072)	(0.0094)	(0.0092)	(0.0115)	(0.0115)
Dominance		0.0100*		0.0069	0.0100*	0.0053	0.0101	0.0061	0.0117	0.0055
		(0.0052)		(0.0051)	(0.0052)	(0.0051)	(0.0067)	(0.0066)	(0.0081)	(0.0080)
Trustworthiness		−0.0086		−0.0107	−0.0079	−0.0092	−0.0111	−0.0103	−0.0032	−0.0072
		(0.0071)		(0.0071)	(0.0071)	(0.0070)	(0.0089)	(0.0088)	(0.0116)	(0.0115)
Human Guess		0.0963***		0.0995***	0.0941***	0.0812***	0.1098***	0.0895***	0.0749*	0.0719*
		(0.0267)		(0.0268)	(0.0268)	(0.0265)	(0.0343)	(0.0339)	(0.0427)	(0.0425)
Socioeconomic Status (SES)	−0.0204***	−0.0162***	−0.0188***	−0.0174***	−0.0153***	−0.0115***				
	(0.0038)	(0.0041)	(0.0039)	(0.0041)	(0.0041)	(0.0040)				
Constant	0.4410***	0.4984***	0.3862***	0.4211***	0.5108***	0.2456***	0.3829***	0.1426***	0.5578***	0.2803***
	(0.0250)	(0.0467)	(0.0241)	(0.0449)	(0.0476)	(0.0447)	(0.0582)	(0.0518)	(0.0770)	(0.0693)
Naive-AUC	0.557	0.604	0.545	0.596	0.604	0.636	0.6	0.647	0.604	0.619
Observations	9,604	9,604	9,604	9,604	9,604	9,604	5,651	5,651	3,953	3,953
Adjusted R^2	0.0072	0.0230	0.0044	0.0200	0.0231	0.0387	0.0194	0.0421	0.0226	0.0300

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features, or the algorithm's overall prediction of judge detention decisions, to actual judge detention decisions, with some control for the defendant's socio-economic status (SES). Specifically we have a separate sample of study subjects independently rate mugshots in the validation (test set) sample for their perceptions of the defendant's SES, then control for that in the regressions shown in columns (1) through (6), which use the full validation (test set) sample. In columns (7) and (8) we re-run the analysis using just those defendants who are above median in their rated SES, while columns (9) and (10) use the remaining sample of defendants.

P-Values: *p<.1; **p<.05; ***p<.01

Table A.XVI: Relationship between novel features and judge decision controlling for defendant's baby-faced feature

	Dependent variable:									
	Judge Detain Decision						BF ≥ Median(BF)		BF < Median(BF)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Face	−0.0213***	−0.0207***			−0.0204***		−0.0170***		−0.0253***	
	(0.0037)	(0.0038)			(0.0038)		(0.0050)		(0.0057)	
Well-Groomed			−0.0170***	−0.0107**	−0.0093*		−0.0042		−0.0155**	
			(0.0043)	(0.0052)	(0.0052)		(0.0069)		(0.0078)	
Algo Judge Detain Prediction						0.6092***		0.6493***		0.5768***
						(0.0437)		(0.0617)		(0.0624)
Male		0.0902***		0.0925***	0.0913***	0.0235**	0.1036***	0.0297*	0.0779***	0.0154
		(0.0108)		(0.0108)	(0.0108)	(0.0117)	(0.0137)	(0.0153)	(0.0176)	(0.0185)
Age		−0.0014***		−0.0017***	−0.0014***	−0.0017***	−0.0012*	−0.0011*	−0.0013**	−0.0019***
		(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Black		−0.0631***		−0.0602***	−0.0630***	−0.0510***	−0.0531***	−0.0364*	−0.0755***	−0.0701***
		(0.0156)		(0.0156)	(0.0156)	(0.0154)	(0.0205)	(0.0203)	(0.0240)	(0.0238)
Asian		−0.0724		−0.0706	−0.0693	−0.0625	−0.0731	−0.0610	−0.0639	−0.0646
		(0.0488)		(0.0489)	(0.0488)	(0.0484)	(0.0641)	(0.0635)	(0.0752)	(0.0746)
Indigenous American		0.0507		0.0688	0.0536	0.0574	0.1277	0.1196	−0.0646	−0.0541
		(0.1019)		(0.1020)	(0.1019)	(0.1010)	(0.1251)	(0.1238)	(0.1745)	(0.1732)
Skin-Tone		−0.1064***		−0.1045***	−0.1078***	−0.0771***	−0.0816**	−0.0503	−0.1379***	−0.1106***
		(0.0250)		(0.0251)	(0.0250)	(0.0249)	(0.0327)	(0.0324)	(0.0389)	(0.0387)
Attractiveness		−0.0080		0.00004	−0.0044	−0.0010	−0.0003	0.0058	−0.0095	−0.0104
		(0.0067)		(0.0070)	(0.0070)	(0.0067)	(0.0093)	(0.0087)	(0.0108)	(0.0103)
Competence		−0.0194***		−0.0178**	−0.0177**	−0.0144**	−0.0181*	−0.0144	−0.0146	−0.0128
		(0.0072)		(0.0073)	(0.0073)	(0.0072)	(0.0098)	(0.0096)	(0.0110)	(0.0108)
Dominance		0.0103**		0.0069	0.0103**	0.0053	0.0132*	0.0077	0.0076	0.0028
		(0.0052)		(0.0051)	(0.0052)	(0.0051)	(0.0070)	(0.0069)	(0.0077)	(0.0076)
Trustworthiness		−0.0079		−0.0091	−0.0070	−0.0085	−0.0064	−0.0075	−0.0102	−0.0115
		(0.0071)		(0.0071)	(0.0071)	(0.0070)	(0.0094)	(0.0093)	(0.0109)	(0.0107)
Human Guess		0.1012***		0.1027***	0.0978***	0.0839***	0.0817**	0.0653*	0.1172***	0.1070***
		(0.0267)		(0.0268)	(0.0268)	(0.0265)	(0.0356)	(0.0352)	(0.0408)	(0.0404)
Baby-Faced (BF)	−0.0108***	−0.0069	−0.0122***	−0.0120***	−0.0061	−0.0066				
	(0.0039)	(0.0043)	(0.0039)	(0.0041)	(0.0043)	(0.0041)				
Constant	0.3902***	0.4709***	0.3645***	0.4215***	0.4892***	0.2339***	0.3636***	0.1195**	0.5714***	0.2987***
	(0.0229)	(0.0477)	(0.0239)	(0.0472)	(0.0488)	(0.0470)	(0.0625)	(0.0549)	(0.0691)	(0.0644)
Naive-AUC	0.547	0.601	0.539	0.595	0.602	0.636	0.602	0.639	0.604	0.631
Observations	9,604	9,604	9,604	9,604	9,604	9,604	5,250	5,250	4,354	4,354
Adjusted R^2	0.0050	0.0217	0.0031	0.0191	0.0219	0.0381	0.0201	0.0383	0.0215	0.0351

Notes: The table presents the results of running separate regressions (one regression per column) that relate the novel facial features, or the algorithm's overall prediction of judge detention decisions, to actual judge detention decisions, with some control for the defendant's perceived baby-facedness. Specifically we have a separate sample of study subjects independently rate mugshots in the validation (test set) sample based on their relative baby-faced looks, and then control for that in the regressions shown in columns (1) through (6), which use the full validation (test set) sample. In columns (7) and (8) we re-run the analysis using just those defendants who are above median in their baby-faced ratings, while columns (9) and (10) use the remaining sample of defendants.

P-Values: *p<.1; **p<.05; ***p<.01

Table A.XVII: Laboratory experiment summary of results

	(1)	(2)	(3)	(4)
Well-Groomed	-0.013*	-0.014**		
	(0.007)	(0.007)		
Heavy-Faced			-0.019***	-0.020***
			(0.007)	(0.007)
Image Pair Fixed Effects?	YES	YES	YES	YES
Participant Fixed Effects?	NO	YES	NO	YES
Number of Subjects	500	500	500	500
Number of Subjects by Image Pair	18,268	18,268	18,548	18,548
Adjusted R^2	0.400	0.401	0.344	0.348

Notes: The table shows the results of two separate randomized lab experiments that randomly morphs pairs of synthetic GAN-generated images in the direction of one of the novel features produced by our hypothesis generation procedure, either well-groomed or heavy-faced; that is, one image within each pair is morphed in the direction of a higher value of the novel feature, and the other image within each pair is morphed in the other direction towards a lower value of the novel feature. We then ask subjects to recommend which of the two defendants they would recommend for detention. Defendants within each pair are also randomly assigned structured variables related to the current charge for which the person was arrested, and their prior criminal record. The table shows the results on the subjectâs detention choice of seeing an image that is more versus less well-groomed (the average difference is 3.7 standard deviations with respect to the distribution of our main GAN-generated mugshot data set) or more versus less heavy-faced (average difference is 4.4 standard deviations). Standard errors are clustered by respondent and image pair. See appendix test for main estimating equation and additional details.

P-Values: *p<0.1; **p<0.05; ***p<0.01

Appendix Figures

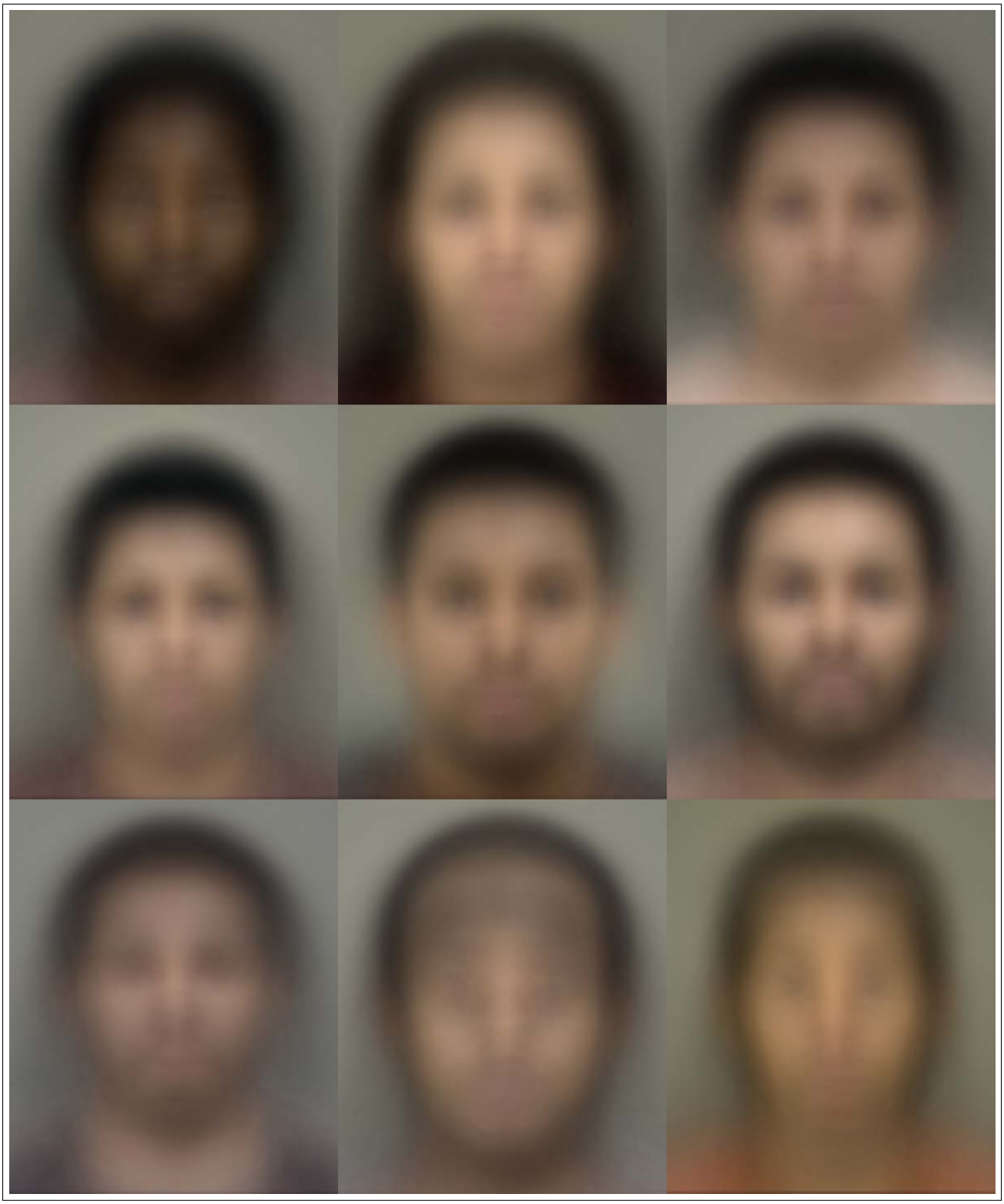


Figure A.I: Eigenfaces

Notes: Eigenfaces method adequately reduces statistical complexity in face image representation but does not provide any interpretable insights for our analysis.

Image number 1.

Questions for image 1

1.

a. Please move the slider to describe how well the face matches each description, from 1 (low) to 9 (high).

Attractiveness: unattractive or unnapealing looks (low) or very attractive (high)

Competence: incompetent appearance (low) or qualified and competent (high)

Dominance: weak or timid (low) or strong and assertive (high)

Trustworthiness: dishonesty (low), or dependable and reliable (high)

Well-groomed Unkempt appearance (low) or well-groomed (high)

Full faced: has gaunt or lean features (low), or chubby, wide set face with broad features (high)

b. Please select the response that you feel best answers the following questions.

What race does this individual appear to be?

- Asian
- Black
- Caucasian / white
- Hispanic
- Indian
- Unsure

What color best matches the natural skin tone of the shown person?

c. Please type your response for the following questions.

What age do you think this individual is? (Use whole years)

Example: 35

Figure A.II: Example of subject labeling exercise for skin-tone, age, and other features

Notes: The mugshot in the above exhibit is a synthetic computer-generated image used for illustration purposes only. In the human intelligence tasks, however, subjects were shown actual defendant mugshots.

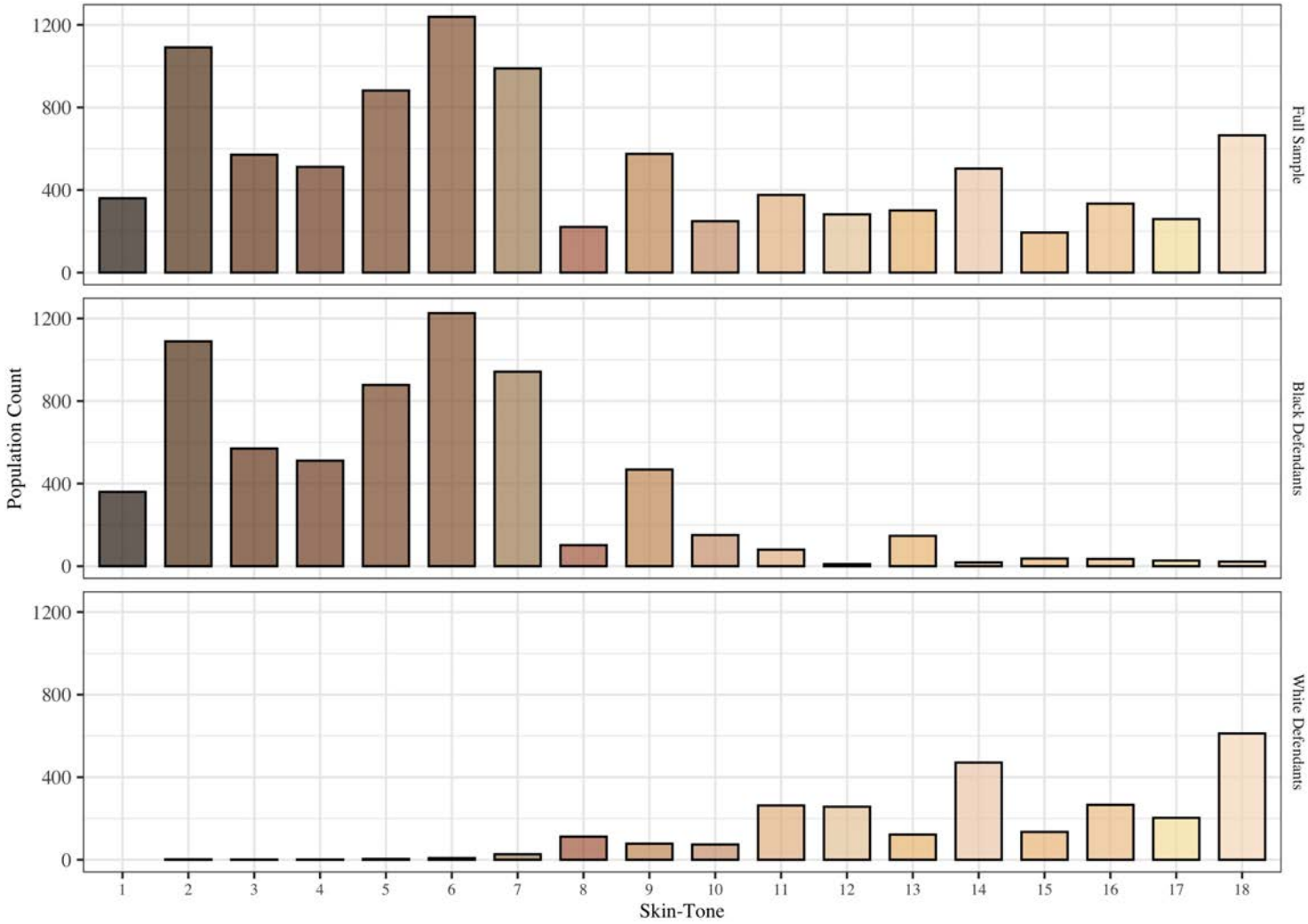
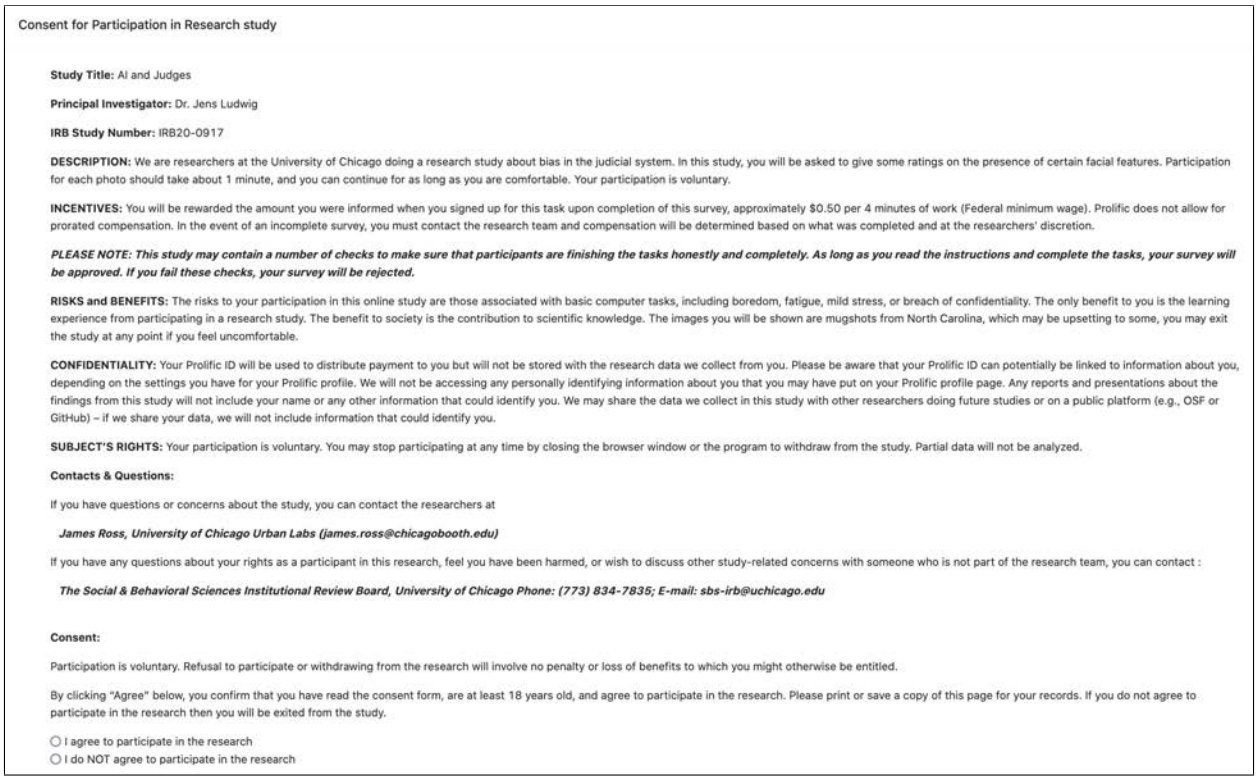


Figure A.III: Distribution of skin-tone categories for full validation sample, and by defendant race

Notes: This figure shows the distribution of skin tone labels from our human intelligence task. These figures come from having human labelers examine face images (mugshots) from Mecklenburg County, NC and recording the skin tone that is closest to the image in the raters view. The top panel shows the histogram of skin tone values reported for the full validation sample; the middle panel is for African American defendants, specifically, while the histogram for white defendants is at the bottom. We collected a total of 10,555 skin tone labels from a total of 77 human raters.



Consent for Participation in Research study

Study Title: AI and Judges

Principal Investigator: Dr. Jens Ludwig

IRB Study Number: IRB20-0917

DESCRIPTION: We are researchers at the University of Chicago doing a research study about bias in the judicial system. In this study, you will be asked to give some ratings on the presence of certain facial features. Participation for each photo should take about 1 minute, and you can continue for as long as you are comfortable. Your participation is voluntary.

INCENTIVES: You will be rewarded the amount you were informed when you signed up for this task upon completion of this survey, approximately $0.50 per 4 minutes of work (Federal minimum wage). Prolific does not allow for prorated compensation. In the event of an incomplete survey, you must contact the research team and compensation will be determined based on what was completed and at the researchers' discretion.

PLEASE NOTE: This study may contain a number of checks to make sure that participants are finishing the tasks honestly and completely. As long as you read the instructions and complete the tasks, your survey will be approved. If you fail these checks, your survey will be rejected.

RISKS and BENEFITS: The risks to your participation in this online study are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. The only benefit to you is the learning experience from participating in a research study. The benefit to society is the contribution to scientific knowledge. The images you will be shown are mugshots from North Carolina, which may be upsetting to some, you may exit the study at any point if you feel uncomfortable.

CONFIDENTIALITY: Your Prolific ID will be used to distribute payment to you but will not be stored with the research data we collect from you. Please be aware that your Prolific ID can potentially be linked to information about you, depending on the settings you have for your Prolific profile. We will not be accessing any personally identifying information about you that you may have put on your Prolific profile page. Any reports and presentations about the findings from this study will not include your name or any other information that could identify you. We may share the data we collect in this study with other researchers doing future studies or on a public platform (e.g., OSF or GitHub) – if we share your data, we will not include information that could identify you.

SUBJECT'S RIGHTS: Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Partial data will not be analyzed.

Contacts & Questions:

If you have questions or concerns about the study, you can contact the researchers at

James Ross, University of Chicago Urban Labs (james.ross@chicagobooth.edu)

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact :

The Social & Behavioral Sciences Institutional Review Board, University of Chicago Phone: (773) 834-7835; E-mail: sbs-irb@uchicago.edu

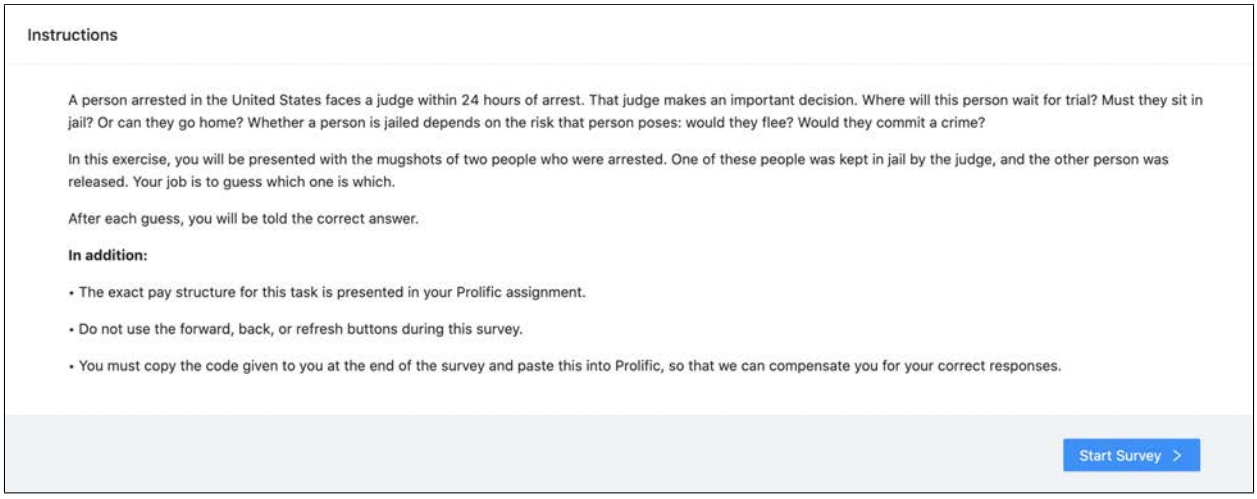
Consent:

Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By clicking "Agree" below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research. Please print or save a copy of this page for your records. If you do not agree to participate in the research then you will be exited from the study.

○ I agree to participate in the research
○ I do NOT agree to participate in the research

(a) The consent screen presented to M-turkers before commencing



Instructions

A person arrested in the United States faces a judge within 24 hours of arrest. That judge makes an important decision. Where will this person wait for trial? Must they sit in jail? Or can they go home? Whether a person is jailed depends on the risk that person poses: would they flee? Would they commit a crime?

In this exercise, you will be presented with the mugshots of two people who were arrested. One of these people was kept in jail by the judge, and the other person was released. Your job is to guess which one is which.

After each guess, you will be told the correct answer.

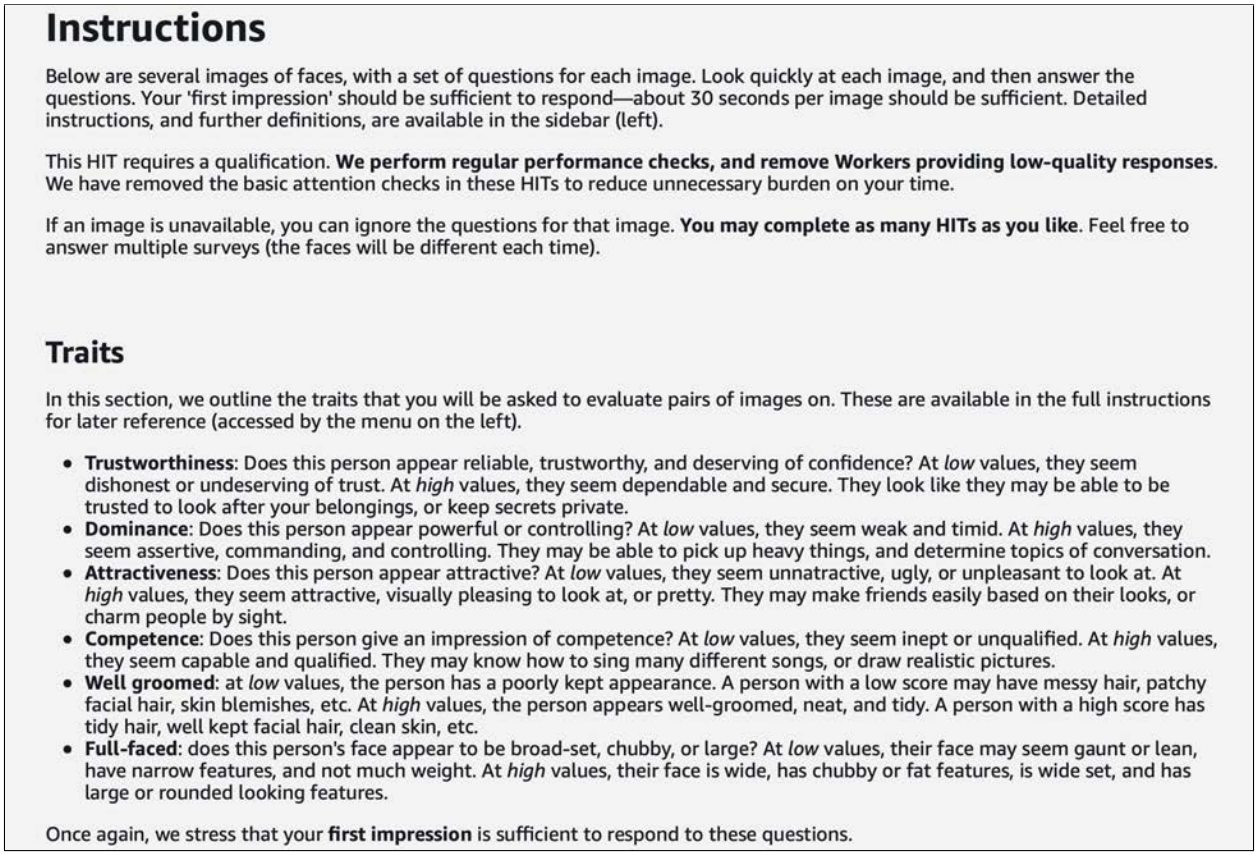
In addition:

- The exact pay structure for this task is presented in your Prolific assignment.
- Do not use the forward, back, or refresh buttons during this survey.
- You must copy the code given to you at the end of the survey and paste this into Prolific, so that we can compensate you for your correct responses.

Start Survey >

(b) The instructions given to Prolific workers for the human guess tasks

Figure A.IV: Examples of consent and instructions shown to M-Turk and Prolific workers for incentivized selection tasks



Instructions

Below are several images of faces, with a set of questions for each image. Look quickly at each image, and then answer the questions. Your 'first impression' should be sufficient to respond—about 30 seconds per image should be sufficient. Detailed instructions, and further definitions, are available in the sidebar (left).

This HIT requires a qualification. **We perform regular performance checks, and remove Workers providing low-quality responses**. We have removed the basic attention checks in these HITs to reduce unnecessary burden on your time.

If an image is unavailable, you can ignore the questions for that image. **You may complete as many HITs as you like**. Feel free to answer multiple surveys (the faces will be different each time).

Traits

In this section, we outline the traits that you will be asked to evaluate pairs of images on. These are available in the full instructions for later reference (accessed by the menu on the left).

- **Trustworthiness**: Does this person appear reliable, trustworthy, and deserving of confidence? At *low* values, they seem dishonest or undeserving of trust. At *high* values, they seem dependable and secure. They look like they may be able to be trusted to look after your belongings, or keep secrets private.
- **Dominance**: Does this person appear powerful or controlling? At *low* values, they seem weak and timid. At *high* values, they seem assertive, commanding, and controlling. They may be able to pick up heavy things, and determine topics of conversation.
- **Attractiveness**: Does this person appear attractive? At *low* values, they seem unnatractive, ugly, or unpleasant to look at. At *high* values, they seem attractive, visually pleasing to look at, or pretty. They may make friends easily based on their looks, or charm people by sight.
- **Competence**: Does this person give an impression of competence? At *low* values, they seem inept or unqualified. At *high* values, they seem capable and qualified. They may know how to sing many different songs, or draw realistic pictures.
- **Well groomed**: at *low* values, the person has a poorly kept appearance. A person with a low score may have messy hair, patchy facial hair, skin blemishes, etc. At *high* values, the person appears well-groomed, neat, and tidy. A person with a high score has tidy hair, well kept facial hair, clean skin, etc.
- **Full-faced**: does this person's face appear to be broad-set, chubby, or large? At *low* values, their face may seem gaunt or lean, have narrow features, and not much weight. At *high* values, their face is wide, has chubby or fat features, is wide set, and has large or rounded looking features.

Once again, we stress that your **first impression** is sufficient to respond to these questions.

Figure A.V: Example of instructions given to M-turkers for one of a labelling task

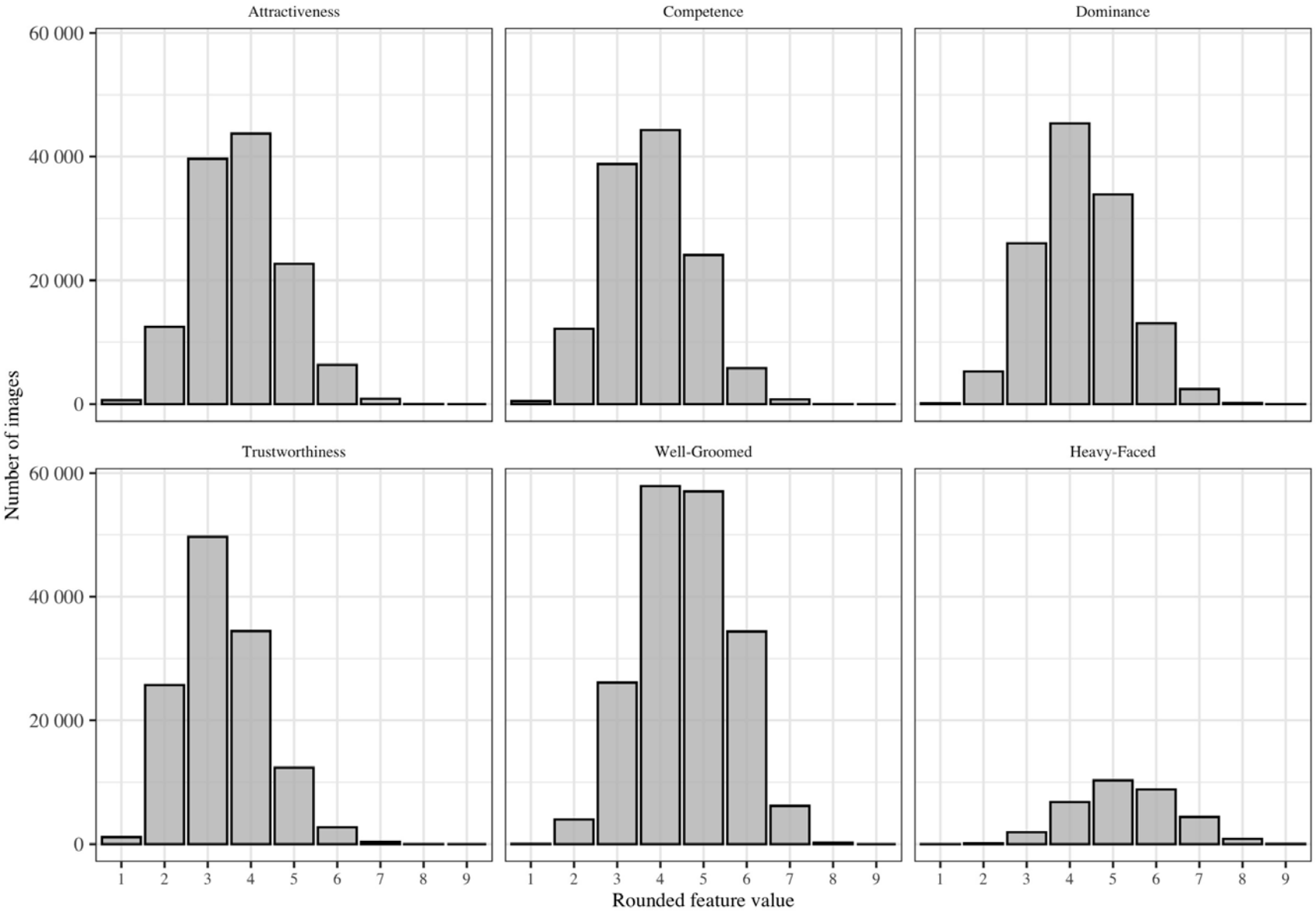


Figure A.VI: Distribution of human ratings of psychological features based on face images

Notes: The standard deviations of these features (calculated on the average label per mugshot) are as follows: attractiveness (0.923), competence (0.911), dominance (0.947), trustworthiness (0.844), well-groomed (1.012), and heavy-faced (1.195).

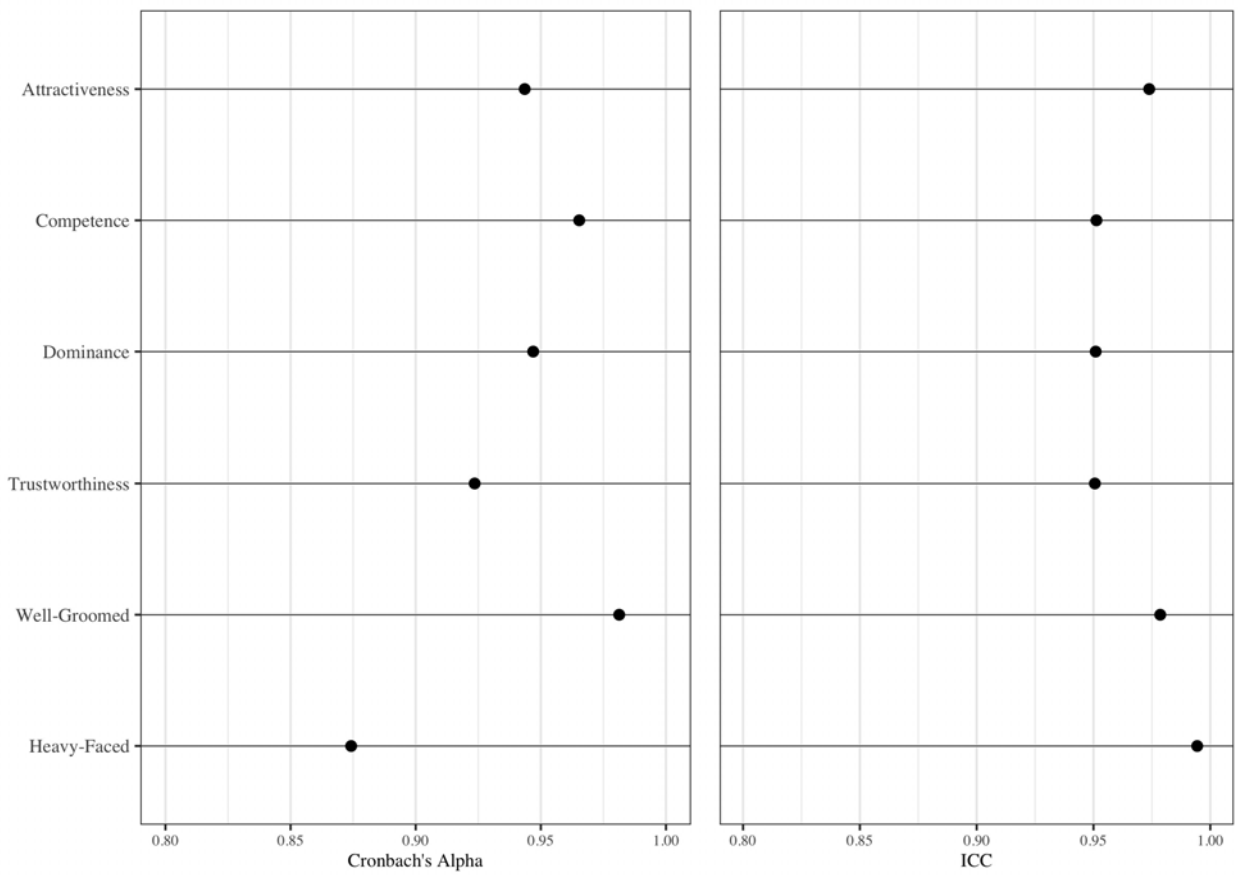


Figure A.VII: Reliability measures for human-rated psychological features

Notes: This figure shows the estimates of Cronbach's alpha (left panel) and Intraclass Correlation Coefficients (right panel) for human ratings of psychological features taken from face images (mugshots) from Mecklenburg County, NC Sheriff's Office public website. Cronbach's alpha (or Tau-equivalent reliability) is a coefficient used to measure the reliability, or internal consistency, of a set of scale or test items. Cronbach's alpha coefficients above 0.80 and 0.90 are considered to be reliable and highly reliable, respectively. Intraclass Correlation Coefficient (ICC) is a continuous inter-rater reliability measure which works for any number of raters giving ratings to a fixed number of items. It provides an estimate of the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings at random. ICC values above 0.80 are considered as an indication of perfect agreement among subjects on the choices of categories. In the above exhibit, Cronbach's alpha coefficients are measured on a bespoke quality check sample while Intraclass Correlation Coefficients are estimated on the entire population of observations.

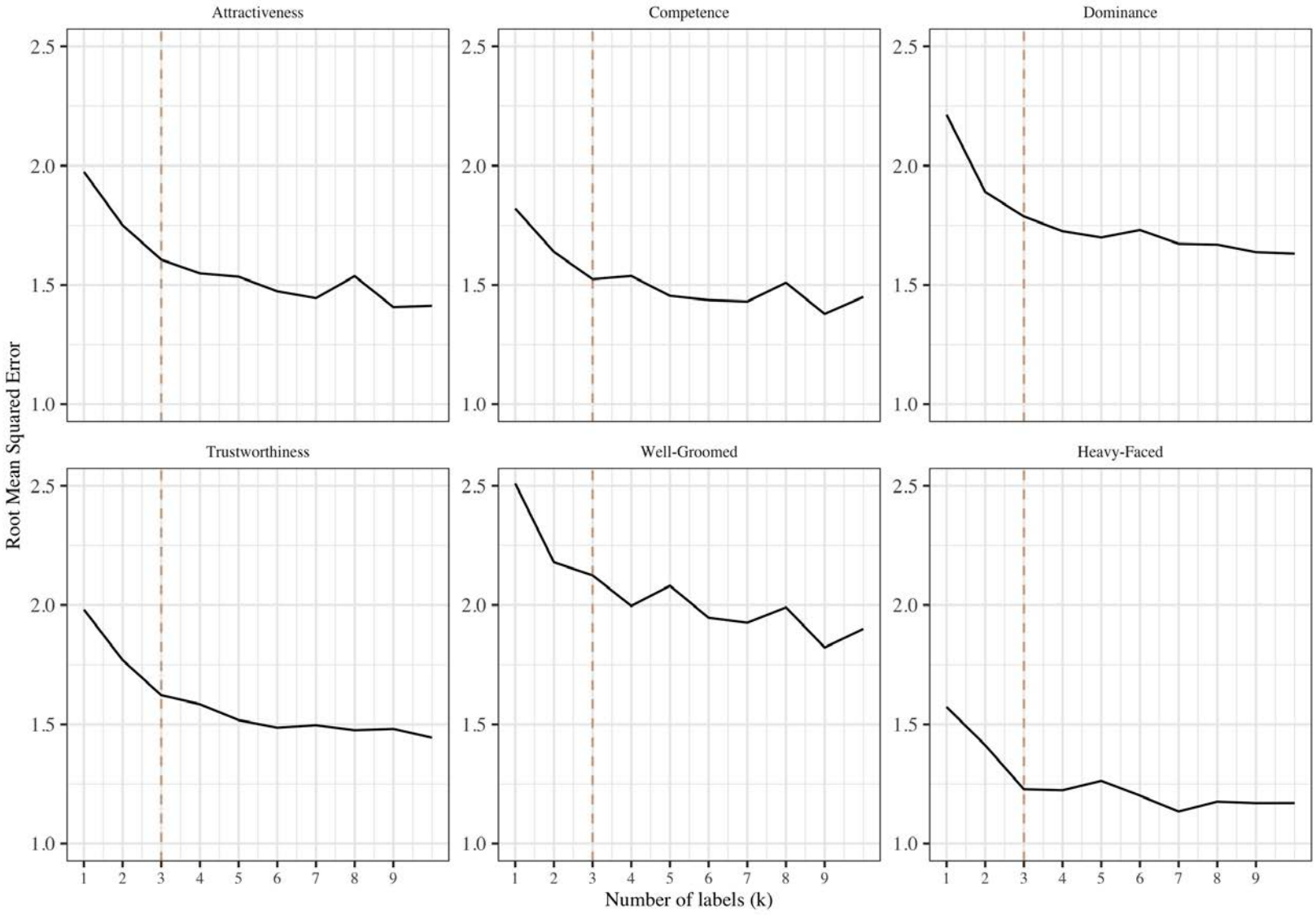
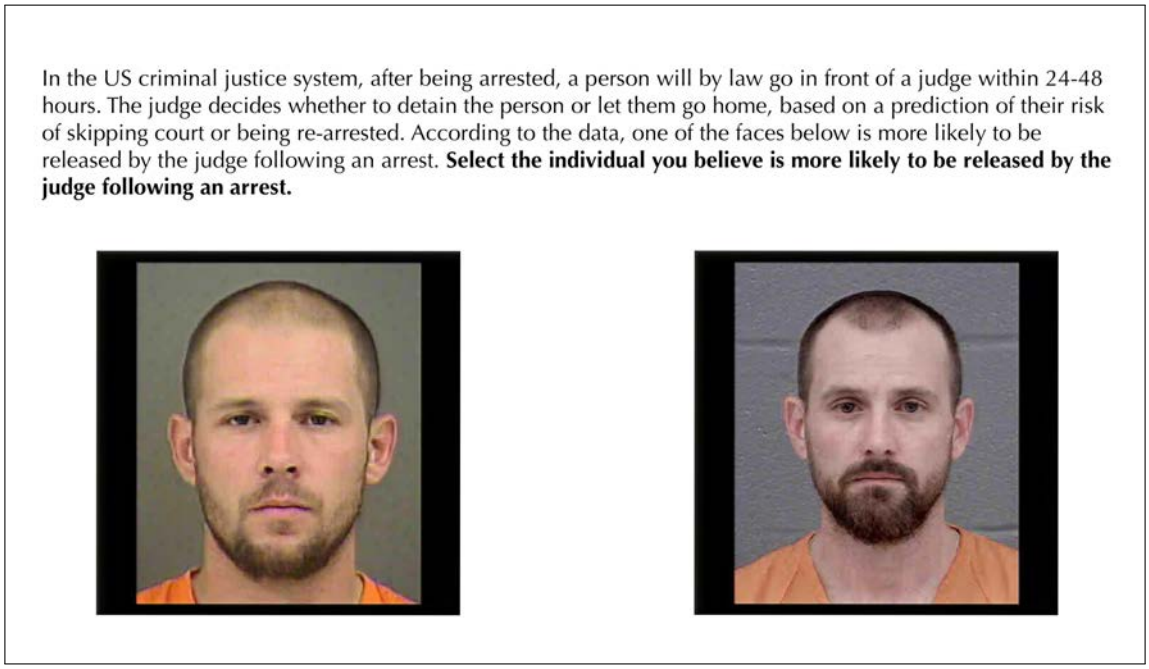
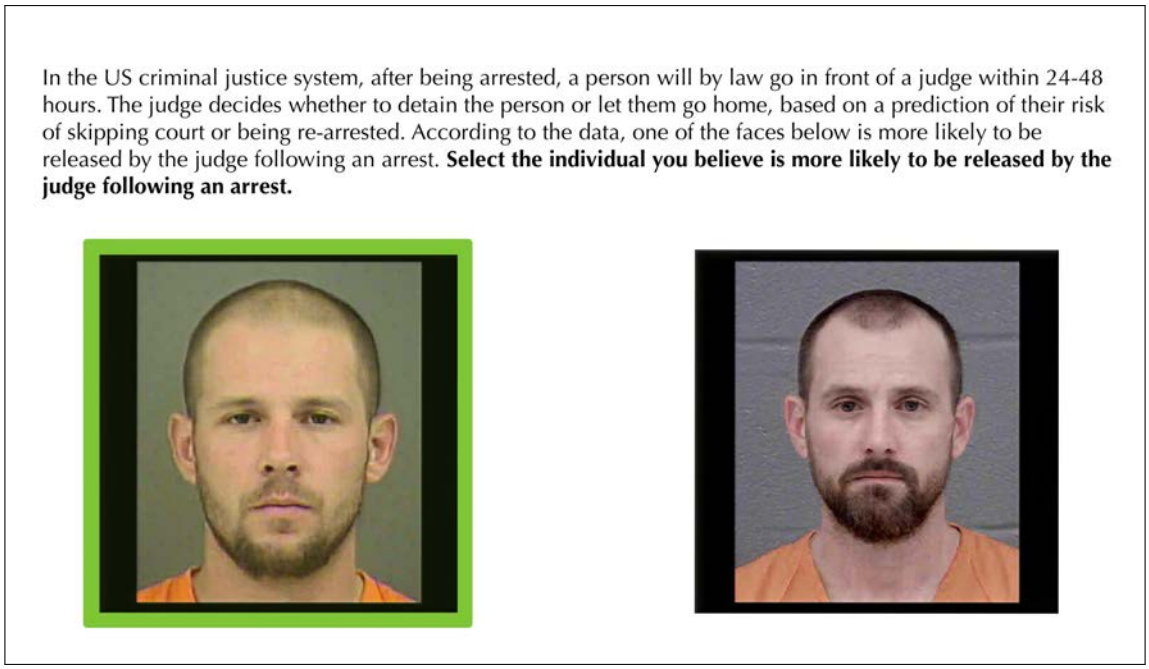


Figure A.VIII: Signal vs. noise in human ratings by number of ratings provided

Notes: The figure shows the results of taking the average of the first K labels provided by human raters for that psychological feature from looking at a face image, and using that to predict the value of the next $(K+1)$ human rating of that same image on the same psychological feature, reported in root mean squared error terms. For each curve relating prediction error and number of labels, we also report the 95% confidence interval.



(a) The screen presented to workers when selecting an image.



(b) The screen presented to workers after selecting an image. In addition to the green outline, a popup window appeared informing candidates if their selection was correct.

Figure A.IX: Example of human intelligence task assessing human performance at picking candidates more likely to be detained.

Notes: The mugshots in the above exhibits are synthetic computer-generated images used for illustration purposes only. In the human intelligence tasks, however, subjects were shown actual defendant mugshots.

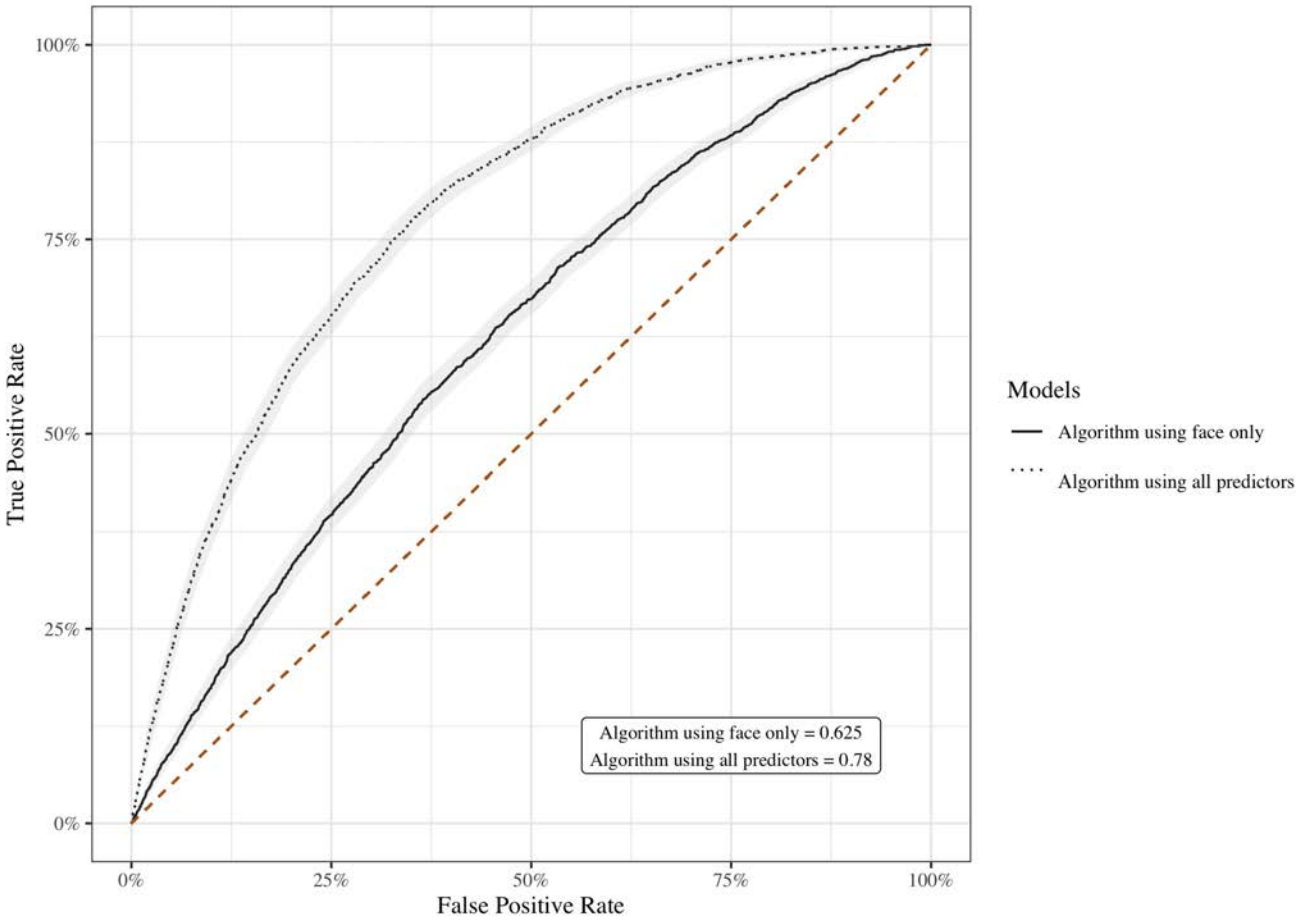


Figure A.X: Accuracy of algorithmic models of judge decisions

Notes: The figure above shows predictive accuracy measures for two separate algorithms built to predict judges' detention decisions, one built using all of the variables available to us from the Mecklenburg County, NC data set (structured variables like current charge, prior record, gender, age, etc.—see text and appendix—as well as unstructured data from defendant's mugshot) and the second built using just the face images alone. The algorithms are built using data from the training data set. We then calculate prediction accuracy out-of-sample on the validation data set (see Table 1 and text). The receiver operating characteristic (ROC) curve plots the true positive rate and false positive rate for all possible classification thresholds; models that are more predictively accurate will have ROC curves that lie relatively further to the northwest. AUC integrates under the ROC curve and can be interpreted as the likelihood that a randomly selected positive (detained) example would be assigned a higher detention likelihood by the algorithm than a randomly selected negative (released) case; random guessing would produce an AUC of 0.5 and perfect prediction would correspond to an AUC of 1.0. The shaded areas correspond to 95% confidence intervals computed using 2,000 stratified bootstrap replicates that sample at the arrestee level.

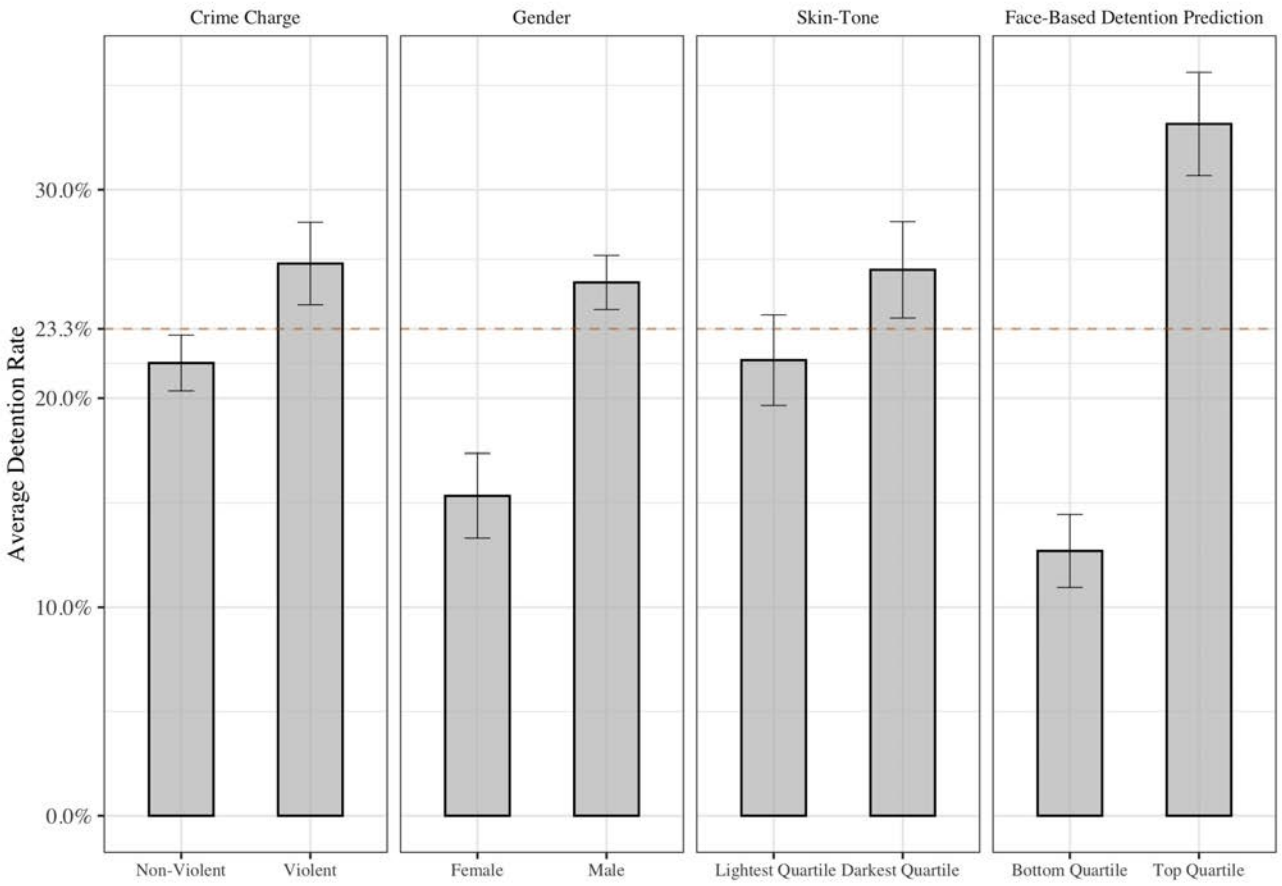
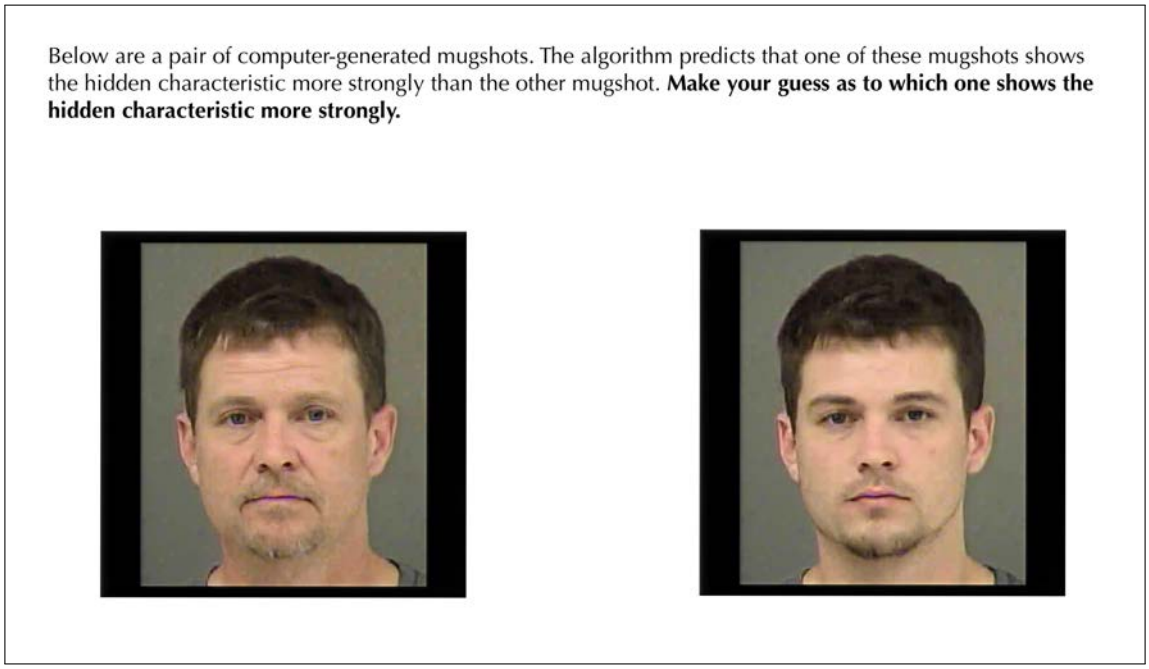
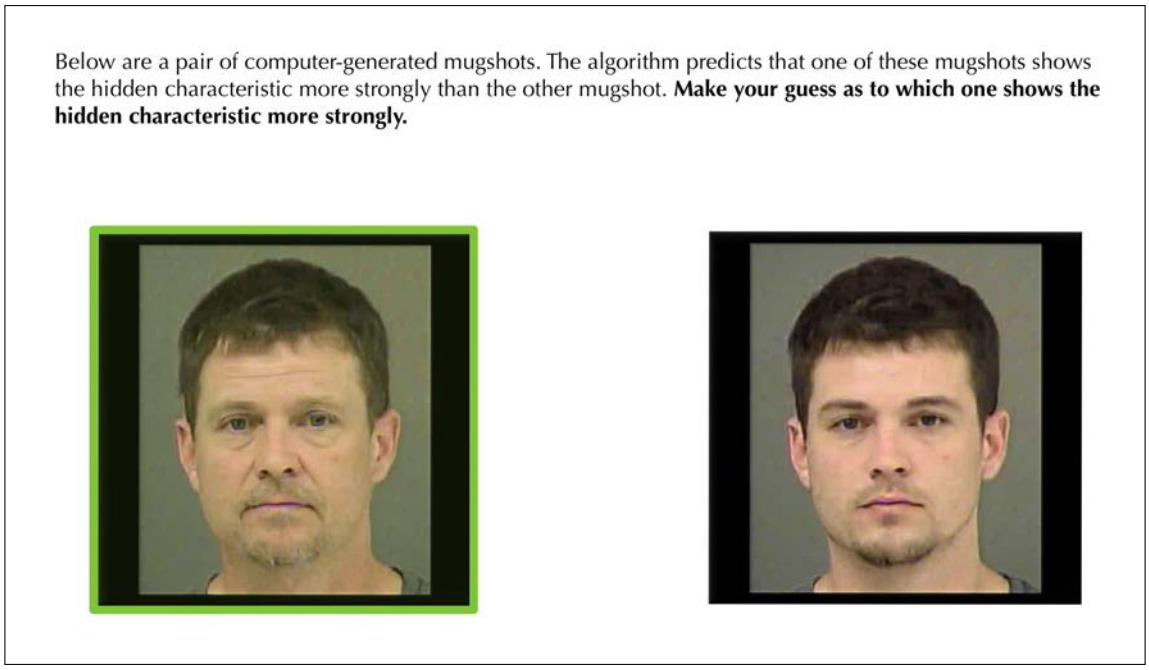


Figure A.XI: Relationship between detention rates and defendant characteristics

Notes: The figure above shows the average validation set detention rates for defendants by different defendant characteristics: crime charge is violent vs. non-violent (first panel), defendant is male versus female (second panel), defendant is in the lightest (Q4) versus darkest (Q1) skin tone shade according to independent subject ratings of mugshots (third panel), and defendant is in lowest quartile of predicted risk (Q1) versus highest quartile (Q4) according to mugshot-based predictor of judge detention decision (final panel). 95% confidence intervals are shown at the top of each bar; overall average detention rate in the validation dataset is 23.3%.



(a) The screen presented to workers when selecting an image.



(b) The screen presented to workers after selecting an image. In addition to the green outline, a popup window appeared informing candidates if their selection was correct.

Figure A.XII: Example of unknown characteristic guessing exercise with predicted-age-morphed pairs

Notes: Subjects were shown age-risk-morphed image pairs and asked to make a guess about the image that exhibited that hidden characteristic more strongly. After completing this guessing exercise on 50 image pairs, subjects were asked to write down the facial features that they believed were related to the algorithm's predictions.

Context: we ran a survey in which several subjects looked at two pictures. One of the pictures was "correct", the other was "incorrect", and the subjects had to guess which was which. After each selection, a popup told them if they were correct or incorrect, and they saw the next pair of photos. We then asked these people to describe how they were selecting the correct answer. That's the data you can see in the Google Doc!

Task: I need you to go through each comment, and "categorize" or "tag" all the comments. You will have to read the comments to discover what categories might exist, and you will have to find every category each comment lies in.

Example: Consider the comment "People with thicker eyebrows were correct, and people who looked energetic, and the ears". There are three different types of categories: a descriptive physical one ('thick eyebrows'), a descriptive impression category ('energetic'), and a vague one ('ears'). We want to tag each of these! The first two are good (this is something specific & measurable), and the last one is bad (not something that can be measured), but we still want the tag.

Challenges: You'll notice they talk about lots of different features, and not always the same ones. Your task: we want to know every different feature mentioned by the subjects, and we want to know how many answers mention each feature. For example, the first response mentioned "a relaxed face". So I went down the entire list of comments, and made a note of every comment that talked about a "relaxed face", or "stressed face", or "relaxed expression", or something similar. The first response also mentioned a "neutral expression", so I went down the list and noted every response that mentioned this, or the opposite. We need to do this for all possible features.

Final state: So, this should be fairly obvious, but our goal is to fill all of the columns with all of the features anybody mentions, and for every feature, we want to note which comments refer to that feature.

Notes:

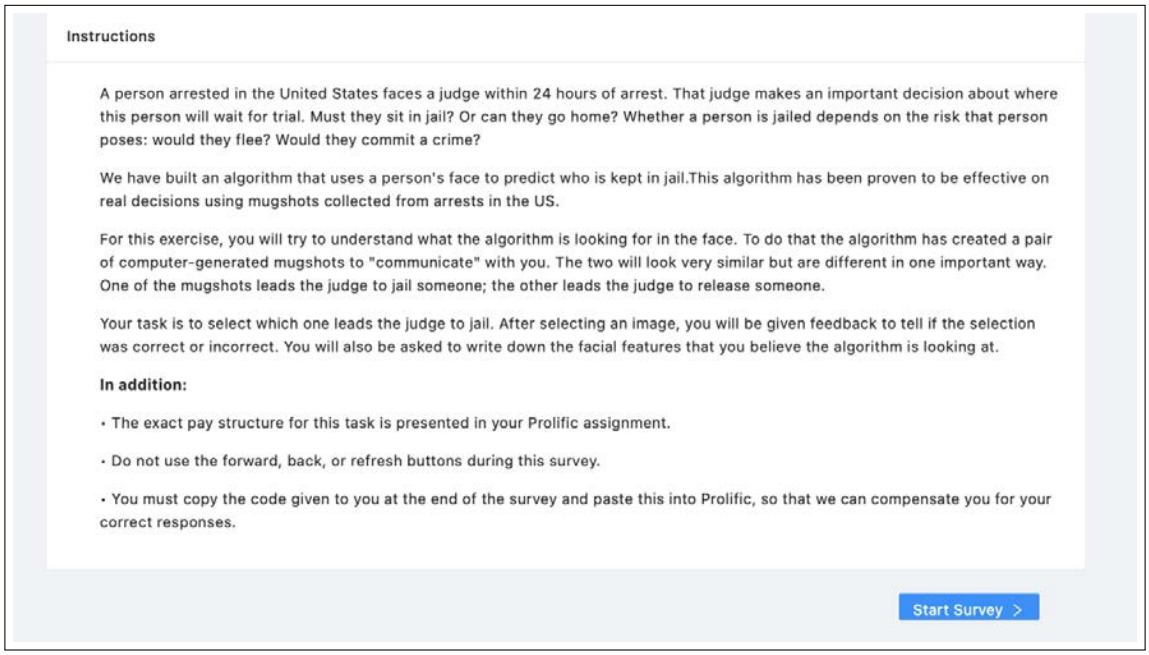
We want to include opposites as the same feature. For example, stressed face / relaxed face is the same feature, since they are opposites; long hair / short hair are the same feature, but not the same feature as curly hair / straight hair; neutral face / happy face are not really the same feature, since the opposite of neutral might be anything.

Features can be something physical (big eyes, crooked nose, long hair) OR something abstract (trustworthy, dangerous looking, competent). Physical features are easy to understand, but abstract features can be complicated. A good rule of thumb here might be: if I asked "based on their face, is this person [trustworthy]?", do you think people would have an answer?

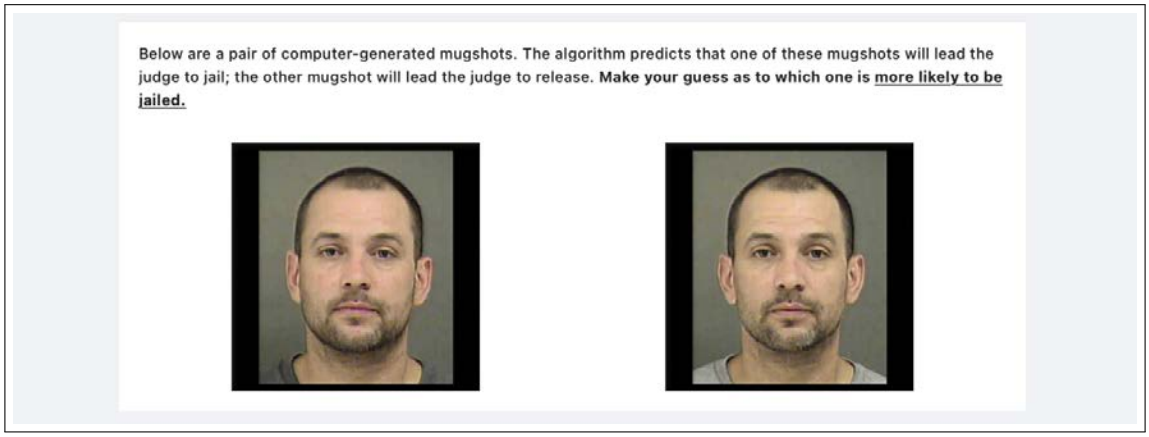
We are looking for features that are specific and measurable. A good rule of thumb is: good features are something about a face, bad features are just parts of a face.

For example, "pursed lips" is good (specific, can be measured as true / false); "looks dangerous" is also good: it's specific (sort of), "short hair" and "long hair" are a single feature; "eyes" is bad (not specific, just a part of a face), so we wouldn't bother tracking this. There are plenty of typos. I think the person who mentioned a bear is really talking about a beard.

Figure A.XIII: Instructions shown to independent RAs for the comment categorization task



(a) Instructions shown to subjects before beginning the task.



(b) The screen presented to workers when selecting an image.

Figure A.XIV: Example of guessing exercise with detention-risk-morphed pairs

Notes: Subjects were shown detention-risk-morphed image pairs such as above and asked to predict which artificial defendant would be more likely to face pre-trial detention. After completing this guessing exercise on 50 image pairs, subjects were asked to write down the facial features that they believed were related to the algorithm's predictions.

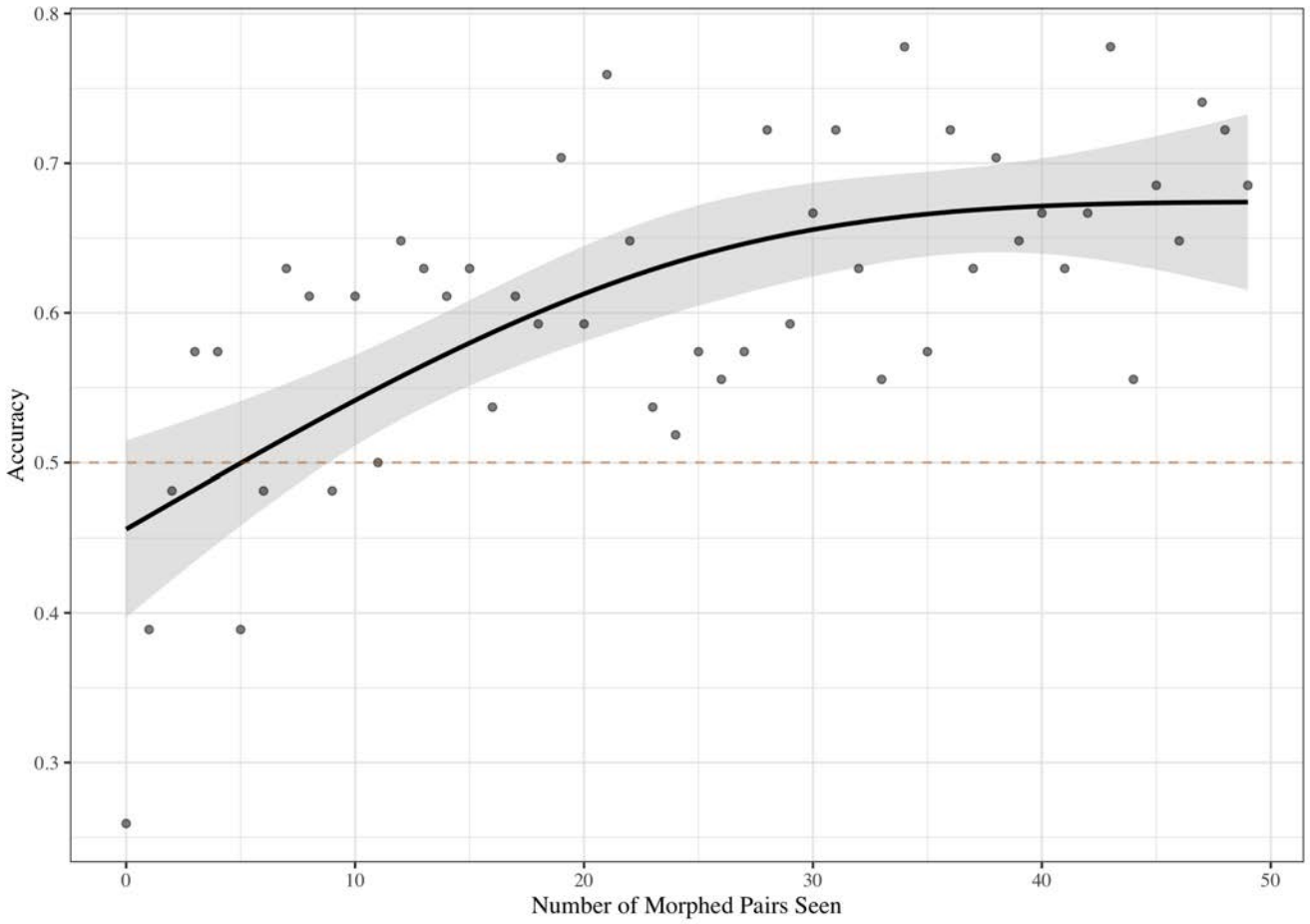


Figure A.XV: Subject performance guessing relative detention risk across morphed image pairs as a function of number of images seen

Notes: The figure above shows subject accuracy rates in guessing which morphed image pair has a higher detention risk, and how that changes as the subjects see more images. Each subject was shown 50 image pairs matched on race, skin tone, age and gender; in our analysis, we treat the data from the first 10 images each subject sees as learning examples and carry out our analyses using the last 40 image-pair results from each subject.

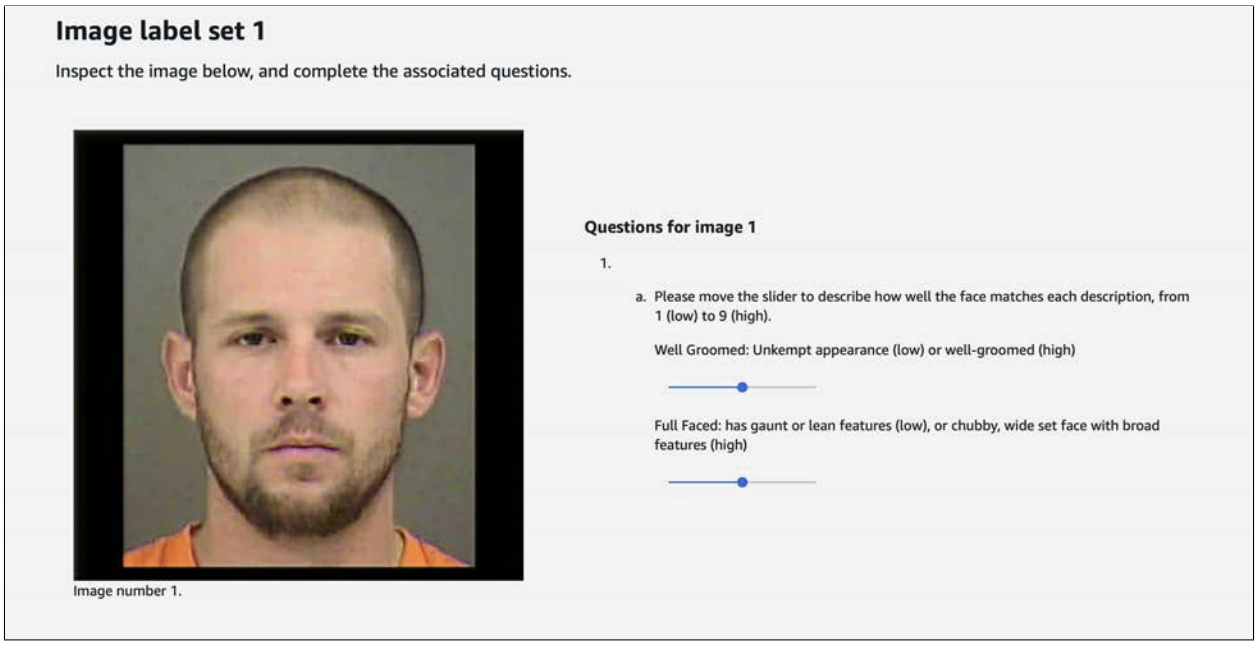


Figure A.XVI: An example of the M-turk labelling exercise

Notes: The mugshot in the above exhibit is a synthetic computer-generated image used for illustration purposes only. In the human intelligence tasks, however, subjects were shown actual defendant mugshots.

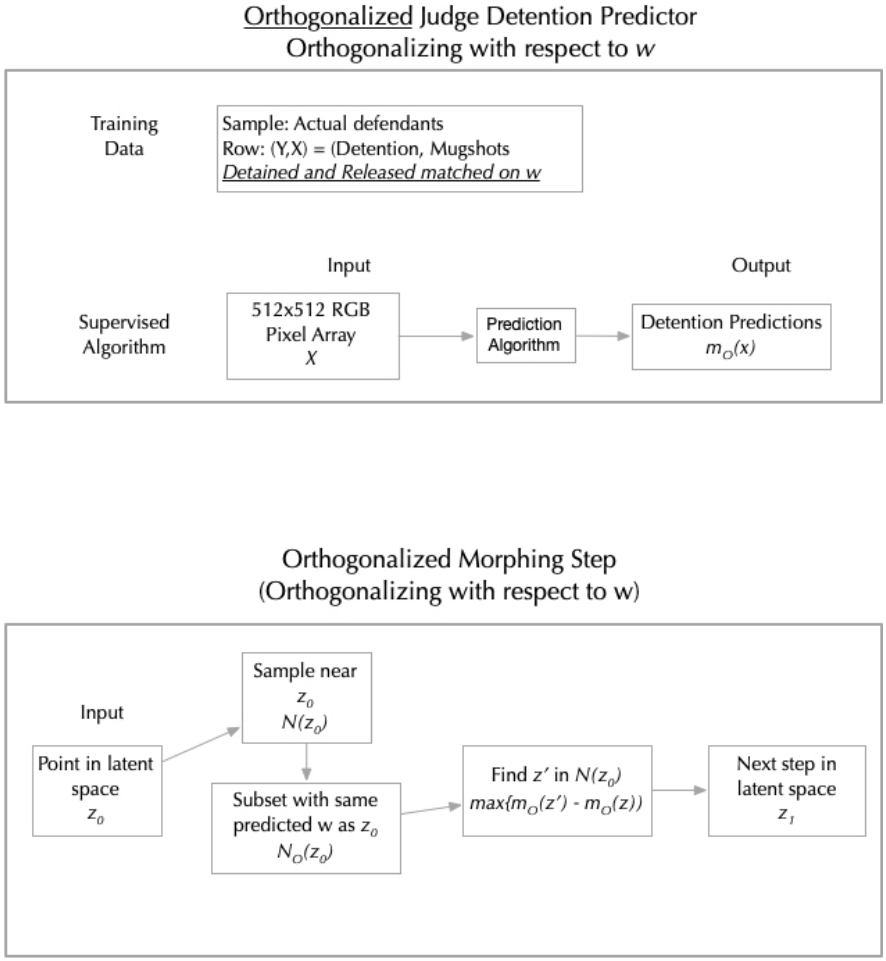
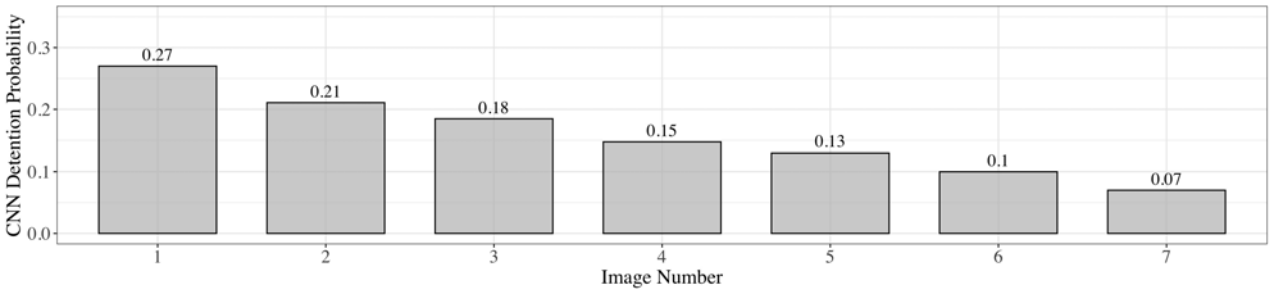


Figure A.XVII: Orthogonalization pipeline

(a) Side-by-side mugshot orthogonal detention morphs with detention probabilities of 0.27 and 0.07 respectively

(b) Transformations of the face along selected steps of the orthogonal morphing process



(c) Detention-probabilities for images in panel (b)

Figure A.XVIII: Illustration of morphed faces along orthogonal gradients of detention predictor

Notes: The top panel shows the result of selecting a random point on the GAN latent face space for a white Hispanic male defendant, then using our orthogonal morphing procedure to increase the predicted detention risk of the image to 0.27 (at left) or reduce the predicted detention risk down to 0.07 (at right); the overall average detention rate in the validation dataset of actual mugshot images is 0.23 by comparison. The second panel shows the different intermediate images between these two end points, while the third panel underneath shows the predicted detention risk for each of the images in the middle panel.

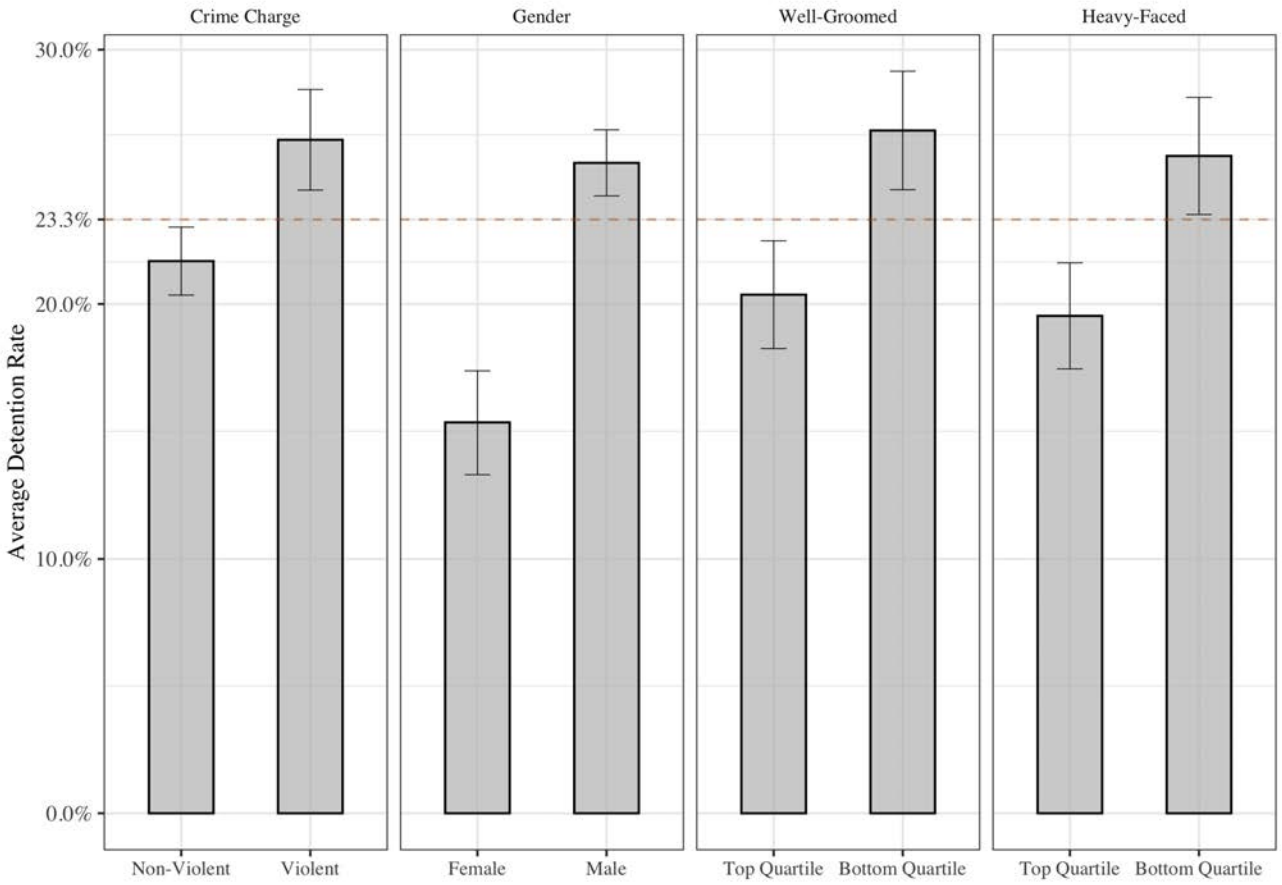
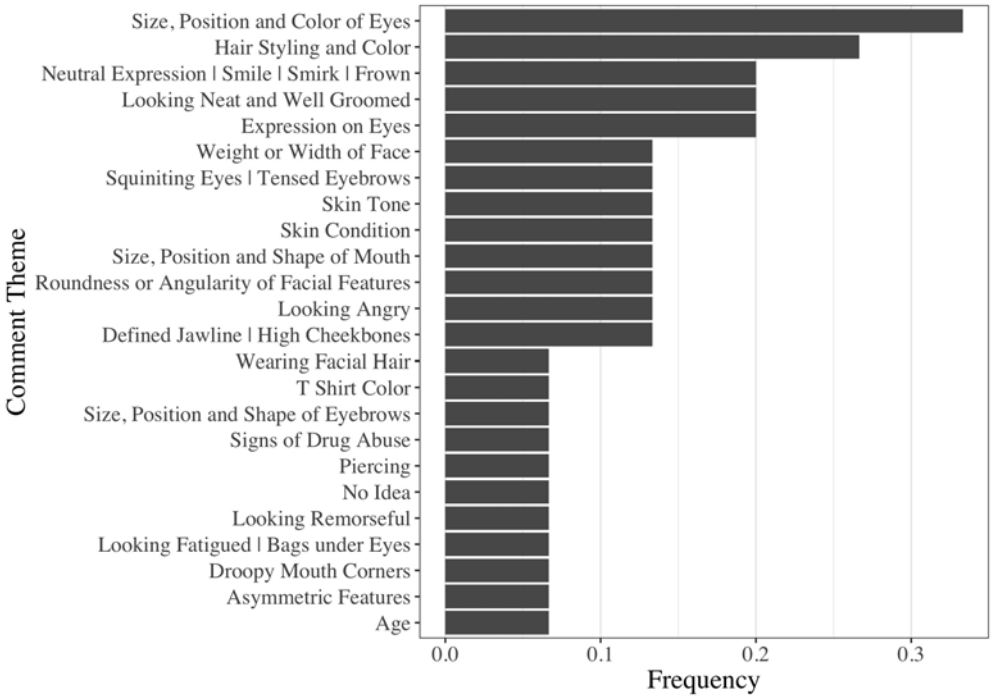


Figure A.XIX: Relative magnitude of the algorithm's discoveries on detention

Notes: The figure above shows the average validation set detention rates among different groups of defendants using charge types, the demographic data of arrestees, and human ratings of our algorithmically generated novel features. The set of bar charts compares the average detention rates for defendants by types of crime charge (violent versus non-violent), by gender (male versus female), and similarly the average detention rates for defendants across top (Q4) and bottom (Q1) quartiles of well-groomed and heavy-faced separately.

(a) A word cloud of practitioners' comments



(b) Frequencies of comments by theme

Figure A.XX: Criminal justice practitioner descriptions of contrast between released and detained actual defendant faces

Notes: The top panel shows a word cloud of subject reports about what they see as the key difference between image pairs, where one is a randomly selected actual mugshot and the other is another actual mugshot which is selected to be congruous in race and gender but discordant in detention outcome. The bottom panel shows the frequency of semantic groupings of these open-ended subject reports (see text for additional detail).

References

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz, "What we teach about race and gender: Representation in images and text of childrenâs books," Technical Report, National Bureau of Economic Research 2021.

Agan, Amanda Y, Jennifer L Doleac, and Anna Harvey, "Misdemeanor prosecution," Technical Report, National Bureau of Economic Research 2021.

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin, "Learning certifiably optimal rule lists for categorical data," *Journal of Machine Learning Research*, 2018, *18*, 1–78.

Angelova, Victoria, Will Dobbie, and Crystal S Yang, "Algorithmic Recommendations and Human Discretion," 2022.

Arnold, David, Will Dobbie, and Crystal S Yang, "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 2018, *133* (4), 1885–1932.

_ , Will S Dobbie, and Peter Hull, "Measuring racial discrimination in bail decisions," Technical Report, National Bureau of Economic Research 2020.

Athey, Susan, "Beyond prediction: Using big data for policy problems," *Science*, 2017, *355* (6324), 483–485.

_ , "The impact of machine learning on economics," in "The economics of artificial intelligence: An agenda," University of Chicago Press, 2018, pp. 507–547.

_ **and Guido W Imbens**, "Machine learning methods that economists should know about," *Annual Review of Economics*, 2019, *11*, 685–725.

_ **, Dean Karlan, Emil Palikot, and Yuan Yuan**, "Smiles in profiles: Improving fairness and efficiency using estimates of user preferences in online marketplaces," Technical Report, National Bureau of Economic Research 2022.

_ **, Guido W Imbens, Jonas Metzger, and Evan Munro**, "Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations," *Journal of Econometrics*, 2021.

Autor, David, "Polanyi's paradox and the shape of employment growth," Technical Report, National Bureau of Economic Research 2014.

Avitzour, Eliana, Adi Choen, Daphna Joel, and Victor Lavy, "On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes," Technical Report, National Bureau of Economic Research 2020.

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, 2010, *11* (1), 1803–1831.

Bai, Xiao, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim, "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," *Pattern Recognition*, 2021, *120*, 108102.

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, 2019, *41* (2), 423–443.

Begall, Sabine, Jaroslav Červenỳ, Julia Neef, Oldřich Vojtěch, and Hynek Burda, "Magnetic alignment in grazing and resting cattle and deer," *Proceedings of the National Academy of Sciences*, 2008, *105* (36), 13451–13455.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*,

2014, *28* (2), 29–50.

Berry, Diane S and Leslie Zebrowitz-McArthur, "What's in a face? Facial maturity and the attribution of legal responsibility," *Personality and Social Psychology Bulletin*, 1988, *14* (1), 23–33.

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American economic review*, 2004, *94* (4), 991–1013.

Bishop, Christopher M and Nasser M Nasrabadi, *Pattern recognition and machine learning*, Vol. 4, Springer, 2006.

Bjornstrom, Eileen ES, Robert L Kaufman, Ruth D Peterson, and Michael D Slater, "Race and ethnic representations of lawbreakers and victims in crime news: A national study of television coverage," *Social problems*, 2010, *57* (2), 269–293.

Breiman, Leo, "Arcing classifier (with discussion and a rejoinder by the author)," *The annals of statistics*, 1998, *26* (3), 801–849.

_ , "Random forests," *Machine learning*, 2001, *45* (1), 5–32.

_ **, Jerome H Friedman, Richard A Olshen, and Charles J Stone**, *Classification and regression trees*, Routledge, 2017.

Brier, Glenn W et al., "Verification of forecasts expressed in terms of probability," *Monthly weather review*, 1950, *78* (1), 1–3.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller, "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*, 2011, *29* (2), 238–249.

Carleo, Giuseppe, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová, "Machine learning and the physical sciences," *Reviews of Modern Physics*, 2019, *91* (4), 045002.

Chang, Chun-Hao, Elliot Creager, Anna Goldenberg, and David Duvenaud, "Explaining image classifiers by counterfactual generation," *arXiv preprint arXiv:1807.08024*, 2018.

Chen, Chaofan and Cynthia Rudin, "An optimization approach to learning falling rule lists," in "International conference on artificial intelligence and statistics" PMLR 2018, pp. 604–612.

_ **, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su**, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, 2019, *32*.

Chen, Daniel L and Arnaud Philippe, "Clash of norms: Judicial leniency on defendant birthdays," *Available at SSRN 3203624*, 2020.

_ **, Tobias J Moskowitz, and Kelly Shue**, "Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires," *The Quarterly Journal of Economics*, 2016, *131* (3), 1181–1242.

Dahl, Gordon B and Matthew M Knepper, "Age discrimination across the business cycle," Technical Report, National Bureau of Economic Research 2020.

Davies, Alex, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász et al., "Advancing mathematics by guiding human intuition with AI," *Nature*, 2021, *600* (7887), 70–74.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Dobbie, Will and Crystal S Yang, "The US pretrial system: Balancing individual rights and

public interests," *Journal of Economic Perspectives*, 2021, *35* (4), 49–70.

_ , **Jacob Goldin, and Crystal S. Yang**, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, February 2018, *108* (2), 201–240.

Doshi-Velez, Finale and Been Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

Eberhardt, Jennifer L, Paul G Davies, Valerie J Purdie-Vaughns, and Sheri Lynn Johnson, "Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes," *Psychological science*, 2006, *17* (5), 383–386.

Einav, Liran and Jonathan Levin, "The data revolution and economic analysis," *Innovation Policy and the Economy*, 2014, *14* (1), 1–24.

Eren, Ozkan and Naci Mocan, "Emotional judges and unlucky juveniles," *American Economic Journal: Applied Economics*, 2018, *10* (3), 171–205.

Freitas, Alex A, "Comprehensible classification models: a position paper," *ACM SIGKDD explorations newsletter*, 2014, *15* (1), 1–10.

Freund, Yoav, Robert Schapire, and Naoki Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, 1999, *14* (5), 771–780.

Frieze, Irene Hanson, Josephine E Olson, and June Russell, "Attractiveness and income for men and women in management," *Journal of Applied Social Psychology*, 1991, *21* (13), 1039–1057.

Fudenberg, Drew and Annie Liang, "Predicting and understanding initial play," *American Economic Review*, 2019, *109* (12), 4112–4141.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–74.

Ghandeharioun, Asma, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W Picard, "Dissect: Disentangled simultaneous explanations via concept traversals," *arXiv preprint arXiv:2105.15164*, 2021.

Ghorbani, Amirata, James Wexler, James Y Zou, and Been Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, 2019, *32.*

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of" blind" auditions on female musicians," *American economic review*, 2000, *90* (4), 715–741.

Goncalves, Felipe and Steven Mello, "A few bad apples? Racial bias in policing," *American Economic Review*, 2021, *111* (5), 1406–1441.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, 2014, *27*, 2672–2680.

_ , _ , _ , _ , _ , _ , _ , **and** _ , "Generative adversarial networks," *Communications of the ACM*, 2020, *63* (11), 139–144.

Grogger, Jeffrey and Greg Ridgeway, "Testing for racial profiling in traffic stops from behind a veil of darkness," *Journal of the American Statistical Association*, 2006, *101* (475), 878–887.

Gurney, Kevin, *An introduction to neural networks*, CRC press, 2018.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer, 2009.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in "Proceedings of the IEEE conference on computer vision and pattern recognition" 2016, pp. 770–778.

He, Siyu, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos, "Learning to predict the cosmological structure formation," *Proceedings of the National Academy of Sciences*, 2019, *116* (28), 13825–13832.

Heckman, James J and Burton Singer, "Abducting economics," *American Economic Review*, 2017, *107* (5), 298–302.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, 2017, *30*.

Heyes, Anthony and Soodeh Saberian, "Temperature and decisions: evidence from 207,000 court cases," *American Economic Journal: Applied Economics*, 2019, *11* (2), 238–265.

Hoekstra, Mark and CarlyWill Sloan, "Does race matter for police use of force? Evidence from 911 calls," *American Economic Review*, 2020, *112* (3), 827–860.

Holte, Robert C, "Very simple classification rules perform well on most commonly used datasets," *Machine learning*, 1993, *11* (1), 63–90.

Hunter, Margaret, "The persistent problem of colorism: Skin tone, status, and inequality," *Sociology compass*, 2007, *1* (1), 237–254.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning*, Vol. 112, Springer, 2013.

Jordan, Michael I and Tom M Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, 2015, *349* (6245), 255–260.

Jr, Roland G Fryer, "An Empirical Analysis of Racial Differences in Police Use of Force: A Response," *Journal of Political Economy*, 2020, *128* (10), 4003–4008.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, 2021, *596* (7873), 583–589.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein, "Simple rules for complex decisions," *arXiv preprint arXiv:1702.04690*, 2017.

Kahneman, Daniel, Olivier Sibony, and CR Sunstein, *Noise*, HarperCollins UK, 2022.

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot, "An adversarial approach to structural estimation," *arXiv preprint arXiv:2007.06169*, 2020.

Karras, Tero, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in "Proceedings of the IEEE/CVF conference on computer vision and pattern recognition" 2019, pp. 4401–4410.

_ , _ , Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition" 2020, pp. 8107–8116.

Kingma, Diederik P and Max Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

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

Korot, Edward, Nikolas Pontikos, Xiaoxuan Liu, Siegfried K Wagner, Livia Faes, Josef Huemer, Konstantinos Balaskas, Alastair K Denniston, Anthony Khawaja, and Pearse A Keane, "Predicting sex from retinal fundus photographs using automated deep learning," *Scientific reports*, 2021, *11* (1), 10286.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012, *25*, 1097–1105.

Lahat, Dana, Tülay Adali, and Christian Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, 2015, *103* (9), 1449–1477.

Lang, Oran, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani et al., "Explaining in style: Training a gan to explain a classifier in stylespace," in "Proceedings of the IEEE/CVF International Conference on Computer Vision" 2021, pp. 693–702.

Langley, Pat, Herbert A Simon, Gary L Bradshaw, and Jan M Zytkow, *Scientific discovery*, Cambridge, Ma: MIT Press, 1987.

LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet, "Convolutional networks and applications in vision," in "Proceedings of 2010 IEEE international symposium on circuits and systems" IEEE 2010, pp. 253–256.

_ , Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, 2015, *521* (7553), 436–444.

Lee, Minhyeok and Junhee Seok, "Controllable generative adversarial network," *Ieee Access*, 2019, *7*, 28158–28169.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg, "Meme-tracking and the dynamics of the news cycle," in "Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining" 2009, pp. 497–506.

Letham, Benjamin, Cynthia Rudin, Tyler H McCormick, and David Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, 2015, *9* (3), 1350–1371.

Li, Oscar, Hao Liu, Chaofan Chen, and Cynthia Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in "Proceedings of the AAAI Conference on Artificial Intelligence," Vol. 32 2018, pp. 3530–3537.

Little, Anthony C, Benedict C Jones, and Lisa M DeBruine, "Facial attractiveness: evolutionary based research," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2011, *366* (1571), 1638–1659.

Liu, Shusen, Bhavya Kailkhura, Donald Loveland, and Yong Han, "Generative counterfactual introspection for explainable deep learning," in "2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)" IEEE 2019, pp. 1–5.

Marcinkevičs, Ričards and Julia E Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv preprint arXiv:2012.01805*, 2020.

Miller, Andrew, Ziad Obermeyer, John Cunningham, and Sendhil Mullainathan, "Discriminative regularization for latent variable models with applications to electrocardiography," in "International Conference on Machine Learning" PMLR 2019, pp. 8072–8081.

Mobius, Markus M and Tanya S Rosenblat, "Why beauty matters," *American Economic Review*, 2006, *96* (1), 222–235.

Mobley, R Keith, *An introduction to predictive maintenance*, Elsevier, 2002.

Mullainathan, Sendhil and Jann Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

_ **and Ziad Obermeyer**, "Diagnosing physician error: A machine learning approach to low-value health care," *The Quarterly Journal of Economics*, 2022, *137* (2), 679–727.

Murphy, Allan H, "A new vector partition of the probability score," *Journal of Applied Meteorology and Climatology*, 1973, *12* (4), 595–600.

Nalisnick, Eric, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan, "Do deep generative models know what they don't know?," *arXiv preprint arXiv:1810.09136*, 2018.

Narayanaswamy, Arunachalam, Subhashini Venugopalan, Dale R Webster, Lily Peng, Greg S Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C Nelson, and Avinash V Varadarajan, "Scientific discovery by generating counterfactuals using image translation," in "International Conference on Medical Image Computing and Computer-Assisted Intervention" Springer 2020, pp. 273–283.

Neumark, David, Ian Burn, and Patrick Button, "Experimental age discrimination evidence and the Heckman critique," *American Economic Review*, 2016, *106* (5), 303–308.

Nielsen, Michael A, *Neural networks and deep learning*, Vol. 25, Determination press San Francisco, CA, 2015.

Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, 2018, *115* (25), E5716–E5725.

Oosterhof, Nikolaas N and Alexander Todorov, "The functional basis of face evaluation," *Proceedings of the National Academy of Sciences*, 2008, *105* (32), 11087–11092.

Peterson, Joshua C, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths, "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, 2021, *372* (6547), 1209–1214.

_ **, Stefan Uddenberg, Thomas L Griffiths, Alexander Todorov, and Jordan W Suchow**, "Deep models of superficial face judgments," *Proceedings of the National Academy of Sciences*, 2022, *119* (17), e2115228119.

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

Pion-Tonachini, Luca, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, Benjamin Nachman et al., "Learning from learning machines: a new generation of AI technology to meet the needs of science," *arXiv preprint arXiv:2111.13786*, 2021.

Popper, Karl, *The logic of scientific discovery*, Routledge, 2005.

Pronin, Emily, "The introspection illusion," *Advances in experimental social psychology*, 2009, *41*, 1–67.

Ramachandram, Dhanesh and Graham W Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, 2017, *34* (6), 96–108.

Rambachan, Ashesh et al., "Identifying prediction mistakes in observational data," *Harvard University*, 2021.

Rawat, Waseem and Zenghui Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, 2017, *29* (9), 2352–2449.

Redcross, Cindy, Britt Henderson, L Miratrix, and E Valentine, "Evaluation of pretrial justice system reforms that use the Public Safety Assessment: Effects in Mecklenburg County

North Carolina Report 2," *MDRC Center for Criminal Justice Research. https://www. mdrc. org/sites/default/files/PSA_Mecklenburg_Brief2. pdf*, 2019.

Rudin, Cynthia, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 2019, *1* (5), 206–215.

_ , Rebecca J Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac, "A process for predicting manhole events in Manhattan," *Machine Learning*, 2010, *80* (1), 1–31.

Said-Metwaly, Sameh, Wim Van den Noortgate, and Eva Kyndt, "Approaches to measuring creativity: A systematic literature review," *Creativity. Theories–Research-Applications*, 2017, *4* (2), 238–275.

Sajjadi, Mehdi SM, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly, "Assessing generative models via precision and recall," *Advances in neural information processing systems*, 2018, *31*.

Schickore, Jutta, "Scientific Discovery," in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, summer 2018 ed., Metaphysics Research Lab, Stanford University, 2018.

Schlag, Pierre, "Law and phrenology," *Harvard Law Review*, 1997, *110* (4), 877–921.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in "In Workshop at International Conference on Learning Representations" Citeseer 2014.

Sirovich, Lawrence and Michael Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a*, 1987, *4* (3), 519–524.

Sunstein, Cass R, "Governing by algorithm? No noise and (potentially) less bias," *Duke LJ*, 2021, *71*, 1175.

Swanson, Don R, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in biology and medicine*, 1986, *30* (1), 7–18.

_ , "Migraine and magnesium: eleven neglected connections," *Perspectives in biology and medicine*, 1988, *31* (4), 526–557.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

Todorov, Alexander and DongWon Oh, "The structure and perceptual basis of social judgments from faces," in "Advances in experimental social psychology," Vol. 63, Elsevier, 2021, pp. 189–245.

_ , Christopher Y Olivola, Ron Dotsch, Peter Mende-Siedlecki et al., "Social attributions from faces: Determinants, consequences, accuracy, and functional significance," *Annual review of psychology*, 2015, *66* (1), 519–545.

Ustun, Berk and Cynthia Rudin, "Learning Optimized Risk Scores," *Journal of Machine Learning Research*, 2019, *20* (150), 1–75.

Varian, Hal R, "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 2014, *28* (2), 3–28.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, 2018, *31* (2), 841–888.

Wilson, Timothy D, *Strangers to ourselves*, Harvard University Press, 2004.

Yegnanarayana, Bayya, *Artificial neural networks*, PHI Learning Pvt. Ltd., 2009.

Yuhas, Ben P, Moise H Goldstein, and Terrence J Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, 1989, *27* (11), 65–71.

Zebrowitz, Leslie A, Victor X Luevano, Philip M Bronstad, and Itzhak Aharon, "Neural activation to babyfaced men matches activation to babies," *Social Neuroscience*, 2009, *4* (1), 1–10.

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu, "Interpretable convolutional neural networks," in "Proceedings of the IEEE conference on computer vision and pattern recognition" 2018, pp. 8827–8836.