MACHINE LEARNING AS A TOOL FOR HYPOTHESIS GENERATION*

Jens Ludwig and Sendhil Mullainathan

While hypothesis testing is a highly formalized activity, hypothesis generation remains largely informal. We propose a systematic procedure to generate novel hypotheses about human behavior, which uses the capacity of machine learning algorithms to notice patterns people might not. We illustrate the procedure with a concrete application: judge decisions about whom to jail. We begin with a striking fact: the defendant's face alone matters greatly for the judge's jailing decision. In fact, an algorithm given only the pixels in the defendant's mug shot accounts for up to half of the predictable variation. We develop a procedure that allows human subjects to interact with this black-box algorithm to produce hypotheses about what in the face influences judge decisions. The procedure generates hypotheses that are both interpretable and novel: they are not explained by demographics (e.g., race) or existing psychology research, nor are they already known (even if tacitly) to people or experts. Though these results are specific, our procedure is general. It provides a way to produce novel, interpretable hypotheses from any high-dimensional data set (e.g., cell phones, satellites, online behavior, news headlines, corporate filings, and high-frequency time series). A central tenet of our article is that hypothesis generation is a valuable activity, and we hope this

*This is a revised version of Chicago Booth working paper 22-15 "Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery." We gratefully acknowledge support from the Alfred P. Sloan Foundation, Emmanuel Roman, and the Center for Applied Artificial Intelligence at the University of Chicago, and we thank Stephen Billings for generously sharing data. For valuable comments we thank Andrei Shleifer, Larry Katz, and five anonymous referees, as well as Marianne Bertrand, Jesse Bruhn, Steven Durlauf, Joel Ferguson, Emma Harrington, Supreet Kaur, Matteo Magnaricotte, Dev Patel, Betsy Levy Paluck, Roberto Rocha, Evan Rose, Suproteem Sarkar, Josh Schwartzstein, Nick Swanson, Nadav Tadelis, Richard Thaler, Alex Todorov, Jenny Wang, and Heather Yang, plus seminar participants at Bocconi, Brown, Columbia, ETH Zurich, Harvard, the London School of Economics, MIT, Stanford, the University of California Berkeley, the University of Chicago, the University of Pennsylvania, the University of Toronto, the 2022 Behavioral Economics Annual Meetings, and the 2022 NBER Summer Institute. For invaluable assistance with the data and analysis we thank Celia Cook, Logan Crowl, Arshia Elyaderani, and especially Jonas Knecht and James Ross. This research was reviewed by the University of Chicago Social and Behavioral Sciences Institutional Review Board (IRB20-0917) and deemed exempt because the project relies on secondary analysis of public data sources. All opinions and any errors are our own.

The Quarterly Journal of Economics (2024), 751–827. https://doi.org/10.1093/qje/qjad055.
Advance Access publication on January 10, 2024.

encourages future work in this largely “prescientific” stage of science. *JEL Codes:* B4, C1.

I. Introduction

Science is curiously asymmetric. New ideas are meticulously tested using data, statistics, and formal models. Yet those ideas originate in a notably less meticulous process involving intuition, inspiration, and creativity. The asymmetry between how ideas are generated versus tested is noteworthy because idea generation is also, at its core, an empirical activity. Creativity begins with “data” (albeit data stored in the mind), which are then “analyzed” (through a purely psychological process of pattern recognition). What feels like inspiration is actually the output of a data analysis run by the human brain. Despite this, idea generation largely happens off stage, something that typically happens before “actual science” begins.[1] Things are likely this way because there is no obvious alternative. The creative process is so human and idiosyncratic that it would seem to resist formalism.

That may be about to change because of two developments. First, human cognition is no longer the only way to notice patterns in the world. Machine learning algorithms can also find patterns, including patterns people might not notice themselves. These algorithms can work not just with structured, tabular data but also with the kinds of inputs that traditionally could only be processed by the mind, like images or text. Second, data on human behavior is exploding: second-by-second price and volume data in asset markets, high-frequency cellphone data on location and usage, CCTV camera and police bodycam footage, news stories, children’s books, the entire text of corporate filings, and so on. The kind of information researchers once relied on for

[1] The question of hypothesis generation has been a vexing one in philosophy, as it appears to follow a process distinct from deduction and has been sometimes called “abduction” (see Schickore 2018 for an overview). A fascinating economic exploration of this topic can be found in Heckman and Singer (2017), which outlines a strategy for how economists should proceed in the face of surprising empirical results. Finally, there is a small but growing literature that uses machine learning in science. In the next section we discuss how our approach is similar in some ways and different in others.

inspiration is now machine readable: what was once solely mental data is increasingly becoming actual data.[2]

We suggest that these changes can be leveraged to expand how hypotheses are generated. Currently, researchers do of course look at data to generate hypotheses, as in exploratory data analysis, but this depends on the idiosyncratic creativity of investigators who must decide what statistics to calculate. In contrast, we suggest capitalizing on the capacity of machine learning algorithms to automatically detect patterns, especially ones people might never have considered. A key challenge is that we require hypotheses that are interpretable to people. One important goal of science is to generalize knowledge to new contexts. Predictive patterns in a single data set alone are rarely useful; they become insightful when they can be generalized. Currently, that generalization is done by people, and people can only generalize things they understand. The predictors produced by machine learning algorithms are, however, notoriously opaque—hard-to-decipher "black boxes." We propose a procedure that integrates these algorithms into a pipeline that results in human-interpretable hypotheses that are both novel and testable.

While our procedure is broadly applicable, we illustrate it in a concrete application: judicial decision making. Specifically we study pretrial decisions about which defendants are jailed versus set free awaiting trial, a decision that by law is supposed to hinge on a prediction of the defendant's risk (Dobbie and Yang 2021).[3] This is also a substantively interesting application in its own right because of the high stakes involved and mounting evidence that judges make these decisions less than perfectly (Kleinberg et al. 2018; Rambachan et al. 2021; Angelova, Dobbie, and Yang 2023).

We begin with a striking fact. When we build a deep learning model of the judge—one that predicts whether the judge will detain a given defendant—a single factor emerges as having large explanatory power: the defendant's face. A predictor that uses only the pixels in the defendant's mug shot explains from one-quarter to nearly one-half of the predictable variation in

[2] See Einav and Levin (2014), Varian (2014), Athey (2017), Mullainathan and Spiess (2017), Gentzkow, Kelly, and Taddy (2019), and Adukia et al. (2023) on how these changes can affect economics.

[3] In practice, there are a number of additional nuances, as discussed in Section III.A and Online Appendix A.A.

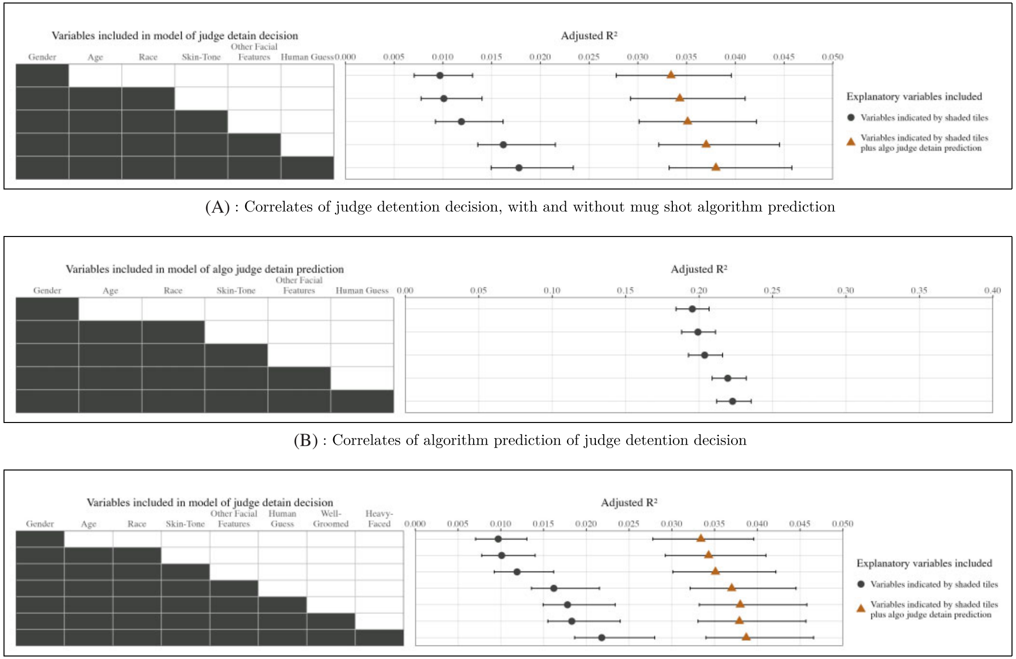
detention.[4] Defendants whose mug shots fall in the bottom quartile of predicted detention are 20.4 percentage points more likely to be jailed than those in the top quartile. By comparison, the difference in detention rates between those arrested for violent versus nonviolent crimes is 4.8 percentage points. Notice what this finding is and is not. We are not claiming the mug shot predicts defendant behavior; that would be the long-discredited field of phrenology (Schlag 1997). We instead claim the mug shot predicts judge behavior: how the defendant looks correlates strongly with whether the judge chooses to jail them.[5]

Has the algorithm found something new in the pixels of the mug shot or simply rediscovered something long known or intuitively understood? After all, psychologists have been studying people's reactions to faces for at least 100 years (Todorov et al. 2015; Todorov and Oh 2021), while economists have shown that judges are influenced by factors (like race) that can be seen from someone's face (Arnold, Dobbie, and Yang 2018; Arnold, Dobbie, and Hull 2020). When we control for age, gender, race, skin color, and even the facial features suggested by previous psychology research (dominance, trustworthiness, attractiveness, and competence), none of these factors (individually or jointly) meaningfully diminishes the algorithm's predictive power (see Figure I, Panel A). It is perhaps worth noting that the algorithm on its own does rediscover some of the signal from these features: in fact, collectively these known features explain 22.3% of the variation in predicted detention (see Figure I, Panel B). The key point is that the algorithm has discovered a great deal more as well.

Perhaps we should control for something else? Figuring out that "something else" is itself a form of hypothesis generation. To avoid a possibly endless—and misleading—process of

[4] This is calculated for some of the most commonly used measures of predictive accuracy, area under the curve (AUC) and R^2, recognizing that different measures could yield somewhat different shares of variation explained. We emphasize the word predictable here: past work has shown that judges are "noisy" and decisions are hard to predict (Kahneman, Sibony, and Sunstein 2022). As a consequence, a predictive model of the judge can do better than the judge themselves (Kleinberg et al. 2018).

[5] In Section IV.B, we examine whether the mug shot's predictive power can be explained by underlying risk differences. There, we tentatively conclude that the predictive power of the face likely reflects judicial error, but that working assumption is not essential to either our results or the ultimate goal of the article: uncovering hypotheses for later careful testing.



(A) : Correlates of judge detention decision, with and without mug shot algorithm prediction

(B) : Correlates of algorithm prediction of judge detention decision

(C) : Correlates of novel features (new hypotheses) and judge detention decision

FIGURE I

Correlates of Judge Detention Decision and Algorithmic Prediction of Judge Decision

FIGURE I

(*Continued*) Panel A summarizes the explanatory power of a regression model in explaining judge detention decisions, controlling for the different explanatory variables indicated at left (shaded tiles), either on their own (dark circles) or together with the algorithmic prediction of the judge decisions (triangles). Each row represents a different regression specification. By "other facial features," we mean variables that previous psychology research suggests matter for how faces influence people's reactions to others (dominance, trustworthiness, competence, and attractiveness). Ninety-five percent confidence intervals around our R^2 estimates come from drawing 10,000 bootstrap samples from the validation data set. Panel B shows the relationship between the different explanatory variables as indicated at left by the shaded tiles with the algorithmic prediction itself as the outcome variable in the regressions. Panel C examines the correlation with judge decisions of the two novel hypotheses generated by our procedure about what facial features affect judge detention decisions: well-groomed and heavy-faced.

generating other controls, we take a different approach. We show mug shots to subjects and ask them to guess whom the judge will detain and incentivize them for accuracy. These guesses summarize the facial features people readily (if implicitly) believe influence jailing. Although subjects are modestly good at this task, the algorithm is much better. It remains highly predictive even after controlling for these guesses. The algorithm seems to have found something novel beyond what scientists have previously hypothesized and beyond whatever patterns people can even recognize in data (whether or not they can articulate them).

What, then, are the novel facial features the algorithm has discovered? If we are unable to answer that question, we will have simply replaced one black box (the judge's mind) with another (an algorithmic model of the judge's mind). We propose a solution whereby the algorithm can communicate what it "sees." Specifically, our procedure begins with a mug shot and "morphs" it to create a mug shot that maximally increases (or decreases) the algorithm's predicted detention probability. The result is pairs of synthetic mug shots that can be examined to understand and articulate what differs within the pairs. The algorithm discovers, and people name that discovery. In principle we could have just shown subjects actual mug shots with higher versus lower predicted detention odds. But faces are so rich that between any pair of actual mug shots, many things will happen to be different and most will be unrelated to detention (akin to the curse of dimensionality). Simply looking at pairs of actual faces can, as a result, lead to many spurious observations. Morphing creates counterfactual synthetic images that are as similar as possible except with respect to detention odds, to minimize extraneous differences and help focus on what truly matters for judge detention decisions.

Importantly, we do not generate hypotheses by looking at the morphs ourselves; instead, they are shown to independent study subjects (MTurk or Prolific workers) in an experimental design. Specifically, we showed pairs of morphed images and asked participants to guess which image the algorithm predicts to have higher detention risk. Subjects were given both incentives and feedback, so they had motivation and opportunity to learn the underlying patterns. While subjects initially guess the judge's decision correctly from these morphed mug shots at about the same rate as they do when looking at "raw data," that is, actual mug shots (modestly above the 50% random guessing mark), they quickly learn from these morphed images what the algorithm is seeing and reach an accuracy of nearly 70%. At the end, participants are asked to put words to the differences they see across images in each pair, that is, to name what they think are the key facial features the algorithm is relying on to predict judge decisions. Comfortingly, there is substantial agreement on what subjects see: a sizable share of subjects all name the same feature. To verify whether the feature they identify is used by the algorithm, a separate sample of subjects independently coded mug shots for this new feature. We show that the new feature is indeed correlated with the algorithm's predictions. What subjects think they're seeing is indeed what the algorithm is also "seeing."

Having discovered a single feature, we can iterate the procedure—the first feature explains only a fraction of what the algorithm has captured, suggesting there are many other factors to be discovered. We again produce morphs, but this time hold the first feature constant: that is, we orthogonalize so that the pairs of morphs do not differ on the first feature. When these new morphs are shown to subjects, they consistently name a second feature, which again correlates with the algorithm's prediction. Both features are quite important. They explain a far larger share of what the algorithm sees than all the other variables (including race and skin color) besides gender. These results establish our main goals: show that the procedure produces meaningful communication, and that it can be iterated.

What are the two discovered features? The first can be called "well-groomed" (e.g., tidy, clean, groomed, versus unkept, disheveled, sloppy look), and the second can be called "heavy-faced" (e.g., wide facial shape, puffier face, wider face, rounder face, heavier). These features are not just predictive of what the algorithm sees, but also of what judges actually do (Figure I, Panel C). We find that both well-groomed and heavy-faced defendants are more likely to be released, even controlling for demographic features and known facial features from psychology. Detention rates of defendants in the top and bottom quartile of well-groomedness differ by 5.5 percentage points (24% of the base rate), while the top versus bottom quartile difference in heavy-facedness is 7 percentage points (about 30% of the base rate). Both differences are larger than the 4.8 percentage points detention rate difference between those arrested for violent versus nonviolent crimes. Not only are these magnitudes substantial, these hypotheses are novel even to practitioners who work in the criminal justice system (in a public defender's office and a legal aid society).

Establishing whether these hypotheses are truly causally related to judge decisions is obviously beyond the scope of the present article. But we nonetheless present a few additional findings that are at least suggestive. These novel features do not appear to be simply proxies for factors like substance abuse, mental health, or socioeconomic status. Moreover, we carried out a lab experiment in which subjects are asked to make hypothetical pretrial release decisions as if they were a judge. They are shown information about criminal records (current charge, prior arrests) along with mug shots that are randomly morphed in the direction of higher or lower values of well-groomed (or heavy-faced). Subjects tend to detain those with higher-risk structured variables (criminal records), all else equal, suggesting they are taking the task seriously. These same subjects, though, are also more likely to detain defendants who are less heavy-faced or well-groomed, even though these were randomly assigned.

Ultimately, though, this is not a study about well-groomed or heavy-faced defendants, nor are its implications limited to faces or judges. It develops a general procedure that can be applied wherever behavior can be predicted using rich (especially high-dimensional) data. Development of such a procedure has required overcoming two key challenges.

First, to generate interpretable hypotheses, we must overcome the notorious black box nature of most machine learning algorithms. Unlike with a regression, one cannot simply inspect the coefficients. A modern deep-learning algorithm, for example, can have tens of millions of parameters. Noninspectability is especially problematic when the data are rich and high dimensional since the parameters are associated with primitives such as pixels. This problem of interpretation is fundamental and remains an active area of research.[6] Part of our procedure here draws on the recent literature in computer science that uses generative models to create counterfactual explanations. Most of those methods are designed for AI applications that seek to automate tasks humans do nearly perfectly, like image classification, where predictability of the outcome (is this image of a dog or a cat?) is typically quite high.[7] Interpretability techniques are used to ensure the algorithm is not picking up on spurious signal.[8] We developed our method, which has similar conceptual underpinnings to this existing literature, for social science applications where the outcome (human behavior) is typically more challenging to predict.[9] To what degree existing methods (as they currently stand or with some modification) could perform as well or better in social science applications like ours is a question we leave to future work.

Second, we must overcome what we might call the Rorschach test problem. Suppose we, the authors, were to look at these morphs and generate a hypothesis. We would not know if the procedure played any meaningful role. Perhaps the morphs, like ink blots, are merely canvases onto which we project our creativity.[10] Put differently, a single research team's idiosyncratic judgments lack the kind of replicability we desire of a scientific procedure. To overcome this problem, it is key that we use independent

[6] For reviews of the interpretability literature, see Doshi-Velez and Kim (2017) and Marcinkevičs and Vogt (2020).

[7] See Liu et al. (2019), Narayanaswamy et al. (2020), Lang et al. (2021), and Ghandeharioun et al. (2022).

[8] For example, if every dog photo in a given training data set had been taken outdoors and every cat photo was taken indoors, the algorithm might learn what animal is in the image based in part on features of the background, which would lead the algorithm to perform poorly in a new data set of more representative images.

[9] For example, for canonical computer science applications like image classification (does this photo contain an image of a dog or of a cat?), predictive accuracy (AUC) can be on the order of 0.99. In contrast, our model of judge decisions using the face only achieves an AUC of 0.625.

[10] Of course even if the hypotheses that are generated are the result of idiosyncratic creativity, this can still be useful. For example, Swanson (1986, 1988) generated two novel medical hypotheses: the possibility that magnesium affects migraines and that fish oil may alleviate Raynaud's syndrome.

(nonresearcher) subjects to inspect the morphs. The fact that a sizable share of subjects all name the same discovery suggests that human-algorithm communication has occurred and the procedure is replicable, rather than reflecting some unique spark of creativity.

At the same time, the fact that our procedure is not fully automatic implies that it will be shaped and constrained by people. Human participants are needed to name the discoveries. So whole new concepts that humans do not yet understand cannot be produced. Such breakthroughs clearly happen (e.g., gravity or probability) but are beyond the scope of procedures like ours. People also play a crucial role in curating the data the algorithm sees. Here, for example, we chose to include mug shots. The creative acquisition of rich data is an important human input into this hypothesis generation procedure.[11]

Our procedure can be applied to a broad range of settings and will be particularly useful for data that are not already intrinsically interpretable. Many data sets contain a few variables that already have clear, fixed meanings and are unlikely to lead to novel discoveries. In contrast, images, text, and time series are rich high-dimensional data with many possible interpretations. Just as there is an ocean of plausible facial features, these sorts of data contain a large set of potential hypotheses that an algorithm can search through. Such data are increasingly available and used by economists, including news headlines, legislative deliberations, annual corporate reports, Federal Open Market Committee statements, Google searches, student essays, résumés, court transcripts, doctors' notes, satellite images, housing photos, and medical images. Our procedure could, for example, raise hypotheses about what kinds of news lead to over- or underreaction of stock prices, which features of a job interview increase racial disparities, or what features of an X-ray drive misdiagnosis.

Central to this work is the belief that hypothesis generation is a valuable activity in and of itself. Beyond whatever the value might be of our specific procedure and empirical application, we hope these results also inspire greater attention to this traditionally "prescientific" stage of science.

[11] Conversely, given a data set, our procedure has a built-in advantage: one could imagine a huge number of hypotheses that, while possible, are not especially useful because they are not measurable. Our procedure is by construction guaranteed to generate hypotheses that are measurable in a data set.

II. A Simple Framework for Discovery

We develop a simple framework to clarify the goals of hypothesis generation and how it differs from testing, how algorithms might help, and how our specific approach to algorithmic hypothesis generation differs from existing methods.[12]

II.A. *The Goals of Hypothesis Generation*

What criteria should we use for assessing hypothesis generation procedures? Two common goals for hypothesis generation are ones that we ensure ex post. First is novelty. In our application, we aim to orthogonalize against known factors, recognizing that it may be hard to orthogonalize against all known hypotheses. Second, we require that hypotheses be testable (Popper 2002). But what can be tested is hard to define ex ante, in part because it depends on the specific hypothesis and the potential experimental setups. Creative empiricists over time often find ways to test hypotheses that previously seemed untestable.[13] To these, we add two more: interpretability and empirical plausibility.

What do we mean by empirically plausible? Let y be some outcome of interest, which for simplicity we assume is binary, and let $h(x)$ be some hypothesis that maps the features of each instance, x, to [0,1]. By empirical plausibility we mean some correlation between y and $h(x)$. Our ultimate aim is to uncover causal relationships. But causality can only be known after causal testing. That raises the question of how to come up with ideas worth causally testing, and how we would recognize them when we see them. Many true hypotheses need not be visible in raw correlations. Those can only be identified with background knowledge (e.g., theory). Other procedures would be required to surface those. Our focus here is on searching for true hypotheses that are visible in raw correlations. Of course not every correlation will turn out to be a true hypothesis, but even in those cases, generating such hypotheses and then invalidating them can be a valuable activity. Debunking spurious correlations has long been one of the most useful roles of empirical work. Understanding what confounders produce those correlations can also be useful.

[12] For additional discussion, see Ludwig and Mullainathan (2023a).

[13] For example, isolating the causal effects of gender on labor market outcomes is a daunting task, but the clever test in Goldin and Rouse (2000) overcomes the identification challenges by using variation in screening of orchestra applicants.

We care about our final goal for hypothesis generation, interpretability, because science is largely about helping people make forecasts into new contexts, and people can only do that with hypotheses they meaningfully understand. Consider an uninterpretable hypothesis like “this set of defendants is more likely to be jailed than that set,” but we cannot articulate a reason why. From that hypothesis, nothing could be said about a new set of courtroom defendants. In contrast an interpretable hypothesis like “skin color affects detention” has implications for other samples of defendants and for entirely different settings. We could ask whether skin color also affects, say, police enforcement choices or whether these effects differ by time of day. By virtue of being interpretable, these hypotheses let us use a wider set of knowledge (police may share racial biases; skin color is not as easily detected at night).[14] Interpretable descriptions let us generalize to novel situations, in addition to being easier to communicate to key stakeholders and lending themselves to interpretable solutions.

II.B. Human versus Algorithmic Hypothesis Generation

Human hypothesis generation has the advantage of generating hypotheses that are interpretable. By construction, the ideas that humans come up with are understandable by humans. But as a procedure for generating new ideas, human creativity has the drawback of often being idiosyncratic and not necessarily replicable. A novel hypothesis is novel exactly because one person noticed it when many others did not. A large body of evidence shows that human judgments have a great deal of “noise.” It is not just that different people draw different conclusions from the same observations, but the same person may notice different things at different times (Kahneman, Sibony, and Sunstein 2022). A large body of psychology research shows that people typically are not able to introspect and understand why we notice specific things those times we do notice them.[15]

[14] See the clever paper by Grogger and Ridgeway (2006) that uses this source of variation to examine this question.

[15] This is related to what Autor (2014) called “Polanyi’s paradox,” the idea that people’s understanding of how the world works is beyond our capacity to explicitly describe it. For discussions in psychology about the difficulty for people to access their own cognition, see Wilson (2004) and Pronin (2009).

There is also no guarantee that human-generated hypotheses need be empirically plausible. The intuition is related to “overfitting.” Suppose that people look at a subset of all data and look for something that differentiates positive ($y = 1$) from negative ($y = 0$) cases. Even with no noise in y, there is randomness in which observations are in the data. That can lead to idiosyncratic differences between $y = 0$ and $y = 1$ cases. As the number of comprehensible hypotheses gets large, there is a “curse of dimensionality”: many plausible hypotheses for these idiosyncratic differences. That is, many different hypotheses can look good in sample but need not work out of sample.[16]

In contrast, supervised learning tools in machine learning are designed to generate predictions in new (out-of-sample) data.[17] That is, algorithms generate hypotheses that are empirically plausible by construction.[18] Moreover, machine learning can detect patterns in data that humans cannot. Algorithms can notice, for example, that livestock all tend to be oriented north (Begall et al. 2008), whether someone is about to have a heart attack based on subtle indications in an electrocardiogram (Mullainathan and Obermeyer 2022), or that a piece of machinery is about to break (Mobley 2002). We call these machine learning prediction functions $m(x)$, which for a binary outcome y map to [0, 1].

[16] Consider a simple example. Suppose $x = (x_1, \ldots, x_k)$ is a k-dimensional binary vector, all possible values of x are equally likely, and the true function in nature relating x to y only depends on the first dimension of x so the function h_1 is the only true hypothesis and the only empirically plausible hypothesis. Even with such a simple true hypothesis, people can generate nonplausible hypotheses. Imagine a pair of data points $(x_0, 0)$ and $(x_1, 1)$. Since the data distribution is uniform, x_0 and x_1 will differ on $\frac{k}{2}$ dimensions in expectation. A person looking at only one pair of observations would have a high chance of generating an empirically implausible hypothesis. Looking at more data, the probability of discovering an implausible hypothesis declines. But the problem remains.

[17] Some canonical references include Breiman et al. (1984), Breiman (2001), Hastie et al. (2009), and Jordan and Mitchell (2015). For discussions about how machine learning connects to economics, see Belloni, Chernozhukov, and Hansen (2014), Varian (2014), Mullainathan and Spiess (2017), Athey (2018), and Athey and Imbens (2019).

[18] Of course there is not always a predictive signal in any given data application. But that is equally an issue for human hypothesis generation. At least with machine learning, we have formal procedures for determining whether there is any signal that holds out of sample.

The challenge is that most $m(x)$ are not interpretable. For this type of statistical model 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 data set) and we used a predictor such as OLS (or LASSO), we could just read the hypotheses from the nonzero coefficients: which variables are significant? Even in that case, interpretation is challenging because machine learning tools, built to generate accurate predictions rather than apportion explanatory power across explanatory variables, yield coefficients that can be unstable across realizations of the data (Mullainathan and Spiess 2017).[19] Often interpretation is much less straightforward than that. If x is an image, text, or time series, the estimated models (such as convolutional neural networks) can have literally millions of parameters. The models are defined on granular inputs with no particular meaning: if we knew $m(x)$ weighted a particular pixel, what have we learned? In these cases, the estimated model $m(x)$ is not interpretable. Our focus is on these contexts where algorithms, as black-box models, are not readily interpreted.

Ideally one might marry people's unique knowledge of what is comprehensible with an algorithm's superior capacity to find meaningful correlations in data: to have the algorithm discover new signal and then have humans name that discovery. How to do so is not straightforward. We might imagine formalizing the set of interpretable prediction functions, and then focus on creating machine learning techniques that search over functions in that set. But mathematically characterizing those functions is typically not possible. Or we might consider seeking insight from a low-dimensional representation of face space, or "eigenfaces," which are a common teaching tool for principal components analysis (Sirovich and Kirby 1987). But those turn out not to provide much useful insight for our purposes.[20] In some sense it

[19] 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 data set, but since the variables are highly correlated, the predicted outcome values themselves (hence predictive accuracy) can be quite stable.

[20] See Online Appendix Figure A.I, which shows the top nine eigenfaces for the data set we describe below, which together explain 62% of the variation.

is obvious why: the subset of actual faces is unlikely to be a linear subspace of the space of pixels. If we took two faces and linearly interpolated them the resulting image would not look like a face. Some other approach is needed. We build on methods in computer science that use generative models to generate counterfactual explanations.

II.C. Related Methods

Our hypothesis generation procedure is part of a growing literature that aims to integrate machine learning into the way science is conducted. A common use (outside of economics) is in what could be called "closed world problems": situations where the fundamental laws are known, but drawing out predictions is computationally hard. For example, the biochemical rules of how proteins fold are known, but it is hard to predict the final shape of a protein. Machine learning has provided fundamental breakthroughs, in effect by making very hard-to-compute outcomes computable in a feasible timeframe.[21]

Progress has been far more limited with applications where the relationship between x and y is unknown ("open world" problems), like human behavior. First, machine learning here has been useful at generating unexpected findings, although these are not hypotheses themselves. Pierson et al. (2021) show that a deep-learning algorithm is better able to predict patient pain from an X-ray than clinicians can: there are physical knee defects that medicine currently does not understand. But that study is not able to isolate what those defects are.[22] Second, machine learning has also been used to explore investigator-generated hypotheses, such as Mullainathan and Obermeyer (2022), who examine whether physicians suffer from limited attention when diagnosing patients.[23]

[21] Examples of applications of this type include Carleo et al. (2019), He et al. (2019), Davies et al. (2021), Jumper et al. (2021), and Pion-Tonachini et al. (2021).

[22] As other examples, researchers have found that retinal images alone can unexpectedly predict gender of patient or macular edema (Narayanaswamy et al. 2020; Korot et al. 2021).

[23] Sheetal, Feng, and Savani (2020) use machine learning to determine which of the long list of other survey variables collected as part of the World Values Survey best predict people's support for unethical behavior. This application sits somewhat in between an investigator-generated hypothesis and the development of an entirely new hypothesis, in the sense that the procedure can only choose

Finally, a few papers take on the same problem that we do. Fudenberg and Liang (2019) and Peterson et al. (2021) have used algorithms to predict play in games and choices between lotteries. They inspected those algorithms to produce their insights. Similarly, Kleinberg et al. (2018) and Sunstein (2021) use algorithmic models of judges and inspect those models to generate hypotheses.[24] Our proposal builds on these papers. Rather than focusing on generating an insight for a specific application, we suggest a procedure that can be broadly used for many applications. Importantly, our procedure does not rely on researcher inspection of algorithmic output. When an expert researcher with a track record of generating scientific ideas uses some procedure to generate an idea, how do we know whether the result is due to the procedure or the researcher? By relying on a fixed algorithmic procedure that human subjects can interface with, hypothesis generation goes from being an idiosyncratic act of individuals to a replicable process.

III. Application and Data

III.A. *Judicial Decision Making*

Although our procedure is broadly applicable, we illustrate it through a specific application to the U.S. criminal justice system. We choose this application partly because of its social relevance. It is also an exemplar of the type of application where our hypothesis generation procedure can be helpful. Its key ingredients—a clear decision maker, a large number of choices (over 10 million people are arrested each year in the United States) that are recorded in data, and, increasingly, high-dimensional data that can also be used to model those choices, such as mug shot images, police body cameras, and text from arrest reports or court transcripts—are shared with a variety of other applications.

Our specific focus is on pretrial hearings. Within 24–48 hours after arrest, a judge must decide where the defendant will await trial, in jail or at home. This is a consequential decision. Cases typically take 2–4 months to resolve, sometimes up to

candidate hypotheses for unethical behavior from the set of variables the World Values Survey investigators thought to include on their questionnaire.

[24] Closest is Miller et al. (2019), which morphs EKG output but stops at the point of generating realistic morphs and does not carry this through to generating interpretable hypotheses.

9–12 months. Jail affects people's families, their livelihoods, and the chances of a guilty plea (Dobbie, Goldin, and Yang 2018). On the other hand, someone who is released could potentially reoffend.[25]

While pretrial decisions are by law supposed to hinge on the defendant's risk of flight or rearrest if released (Dobbie and Yang 2021), studies show that judges' decisions deviate from those guidelines in a number of ways. For starters, judges seem to systematically mispredict defendant risk (Jung et al. 2017; Kleinberg et al. 2018; Rambachan 2021; Angelova, Dobbie, and Yang 2023), partly because judges overweight the charge for which people are arrested (Sunstein 2021). Judge decisions can also depend on extralegal factors like race (Arnold, Dobbie, and Yang 2018; Arnold, Dobbie, and Hull 2020), whether the judge's favorite football team lost (Eren and Mocan 2018), weather (Heyes and Saberian 2019), the cases the judge just heard (Chen, Moskowitz, and Shue 2016), and if the hearing is on the defendant's birthday (Chen and Philippe 2023). These studies test hypotheses that some human being was clever enough to think up. But there remains a great deal of unexplained variation in judges' decisions. The challenge of expanding the set of hypotheses for understanding this variation without losing the benefit of interpretability is the motivation for our own analysis here.

III.B. Administrative Data

We obtained data from Mecklenburg County, North Carolina, the second most populated county in the state (over 1 million residents) that includes North Carolina's largest city (Charlotte). The county is similar to the rest of the United States in terms of economic conditions (2021 poverty rates were 11.0% versus 11.4%, respectively), although the share of Mecklenburg County's population that is non-Hispanic white is lower than the United States as a whole (56.6% versus 75.8%).[26] We rely on three sources of administrative data:[27]

[25] Additional details about how the system works are found in Online Appendix A.

[26] For Black non-Hispanics, the figures for Mecklenburg County versus the United States were 33.3% versus 13.6%. See https://www.census.gov/programs-surveys/sis/resources/data-tools/quickfacts.html.

[27] Details on how we operationalize these variables are found in Online Appendix A.

- The Mecklenburg County Sheriff's Office (MCSO) publicly posts arrest data for the past three years, which provides information on defendant demographics like age, gender, and race, as well as the charge for which someone was arrested.
- The North Carolina Administrative Office of the Courts (NCAOC) maintains records on the judge's pretrial decisions (detain, release, etc.).
- Data from the North Carolina Department of Public Safety includes information about the defendant's prior convictions and incarceration spells, if any.

We also downloaded photos of the defendants from the MCSO public website (so-called mug shots),[28] which capture a frontal view of each person from the shoulders up in front of a gray background. These images are 400 pixels wide by 480 pixels high, but we pad them with a black boundary to be square 512 × 512 images to conform with the requirements of some of the machine learning tools. In Figure II, we give readers a sense of what these mug shots look like, with two important caveats. First, given concerns about how the overrepresentation of disadvantaged groups in discussions of crime can contribute to stereotyping (Bjornstrom et al. 2010), we illustrate the key ideas of the paper using images for non-Hispanic white males. Second, out of sensitivity to actual arrestees, we do not wish to display actual mug shots (which are available at the MCSO website).[29] Instead, the article only shows mug shots that are synthetic, generated using generative adversarial networks as described in Section V.B.

These data capture much of the information the judge has available at the time of the pretrial hearing, but not all of it. Both the judge and the algorithm see structured variables about each defendant like defendant demographics, current charge, and prior record. Because the mug shot (which the algorithm uses) is taken not long before the pretrial hearing, it should be a reasonable proxy for what the judge sees in court. The additional information the judge has but the algorithm does not includes the narrative

[28] The mug shot seems to have originated in Paris in the 1800s (https://law.marquette.edu/facultyblog/2013/10/a-history-of-the-mug-shot/). The etymology of the term is unclear, possibly based on "mug" as slang for either the face or an "incompetent person" or "sucker" since only those who get caught are photographed by police (https://www.etymonline.com/word/mug-shot).

[29] See https://mecksheriffweb.mecklenburgcountync.gov/.

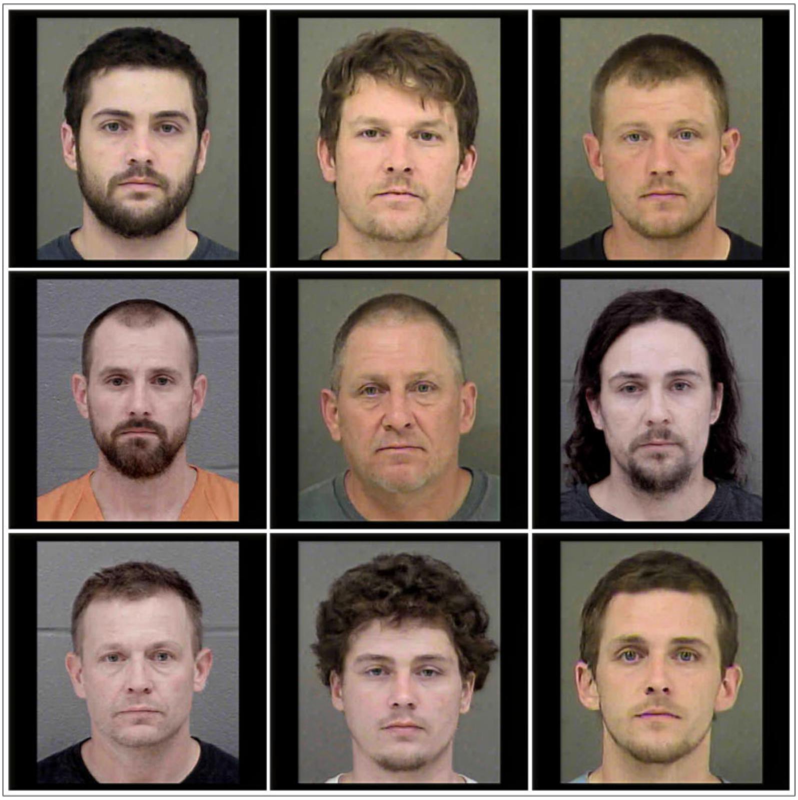


FIGURE II

Illustrative Facial Images

This figure shows facial images that illustrate the format of the mug shots posted publicly on the Mecklenberg County, North Carolina, sheriff's office website. These are not real mug shots of actual people who have been arrested, but are synthetic. Moreover, given concerns about how the overrepresentation of disadvantaged groups in discussions of crime can exacerbate stereotyping, we illustrate the our key ideas using images for non-Hispanic white men. However, in our human intelligence tasks that ask participants to provide labels (ratings for different image features), we show images that are representative of the Mecklenberg County defendant population as a whole.

arrest report from the police and what happens in court. While pretrial hearings can be quite brief in many jurisdictions (often not more than just a few minutes), the judge may nonetheless hear statements from police, prosecutors, defense lawyers, and sometimes family members. Defendants usually have their lawyers speak for them and do not say much at these hearings.

We downloaded 81,166 arrests made between January 18, 2017, and January 17, 2020, involving 42,353 unique defendants. We apply several data filters, like dropping cases without mugshots (Online Appendix Table A.I), leaving 51,751 observations. Because our goal is inference about new out-of-sample (OOS) observations, we partition our data as follows:

- A train set of N = 22,696 cases, constructed by taking arrests through July 17, 2019, grouping arrests by arrestee,[30] randomly selecting 70% to the training-plus-validation data set, then randomly selecting 70% of those arrestees for the training data specifically.
- A validation set of N = 9,604 cases used to report OOS performance in the article's main exhibits, consisting of the remaining 30% in the combined training-plus-validation data frame.
- A lock-box hold-out set of N = 19,009 cases that we did not touch until the article was accepted for final publication, to avoid what one might call researcher overfitting: we run lots of models over the course of writing the article, and the results on the validation data set may overstate our findings. This data set consists of the N = 4,759 valid cases for the last six months of our data period (July 17, 2019, to January 17, 2020) plus a random sample of 30% of those arrested before July 17, 2019, so that we can present results that are OOS with respect to individuals and time. Once this article was officially accepted, we replicated the findings presented in our main exhibits (see Online Appendix D and Online Appendix Tables A.XVIII–A.XXXII). We see that our core findings are qualitatively similar.[31]

Descriptive statistics are shown in Table I. Relative to the county as a whole, the arrested population substantially

[30] We partition the data by arrestee, not arrest, to ensure people show up in only one of the partitions to avoid inadvertent information "leakage" across data partitions.

[31] As the Online Appendix tables show, while there are some changes to a few of the coefficients that relate the algorithm's predictions to factors known from past research to shape human decisions, the core findings and conclusions about the importance of the defendant's appearance and the two specific novel facial features we identify are similar.

TABLE I

SUMMARY STATISTICS FOR MECKLENBURG COUNTY NC DATA, 2017–2020

	Train + validation set	Train set	Validation set	Complete lock-box hold-out	Lock-box hold-out data (OOS by individual)	Lock-box hold-out data (OOS by time)
Sample size	32,300	22,696	9,604	19,009	14,250	4,759
Outcome						
Judge detains defendant	0.233	0.232	0.233	0.214	0.235	0.152
Defendant rearrested before trial	0.251	0.251	0.251	0.202	0.255	0.043
Defendant characteristics						
Age	31.785	31.849	31.631	32.439	32.171	33.239
Male	0.787	0.789	0.782	0.770	0.778	0.747
White	0.277	0.278	0.274	0.295	0.285	0.324
Black	0.694	0.694	0.695	0.677	0.687	0.647
Other race	0.029	0.028	0.031	0.027	0.027	0.029
Arrest year						
2017	0.359	0.359	0.358	0.267	0.357	0.000
2018	0.411	0.411	0.412	0.312	0.416	0.000
2019	0.230	0.230	0.230	0.420	0.228	0.996
Arrest charge						
Violent	0.343	0.343	0.343	0.345	0.339	0.364
Property	0.322	0.324	0.317	0.311	0.319	0.284
Drug	0.205	0.204	0.207	0.186	0.198	0.148
Gun	0.081	0.079	0.084	0.077	0.078	0.072
Other	0.263	0.262	0.264	0.278	0.272	0.294

TABLE I
CONTINUED

	Train + validation set	Train set	Validation set	Complete lock-box hold-out	Lock-box hold-out data (OOS by individual)	Lock-box hold-out data (OOS by time)
Arrest charge severity						
Felony	0.423	0.421	0.428	0.400	0.410	0.370
Non-felony	0.577	0.579	0.572	0.600	0.590	0.630
Defendant prior record						
Any prior conviction	0.460	0.461	0.458	0.425	0.452	0.344
Prior felony conviction	0.331	0.333	0.328	0.302	0.323	0.240
Prior non-felony conviction	0.316	0.316	0.318	0.296	0.313	0.244

Notes. This table reports descriptive statistics for our full data set and analysis subsets, which cover the period January 18, 2017, through January 17, 2020, from Mecklenburg County, NC. The lock-box hold-out data set consists of data from the last six months of our study period (July 17, 2019–January 17, 2020) 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 to our validation set (which we use to report results in the article's main exhibits). For additional details of our data filters and partitioning procedures, see Online Appendix Table A.I. We define pretrial release as being released on the defendant's own recognizance or having been assigned and then posting cash bail requirements within three days of arrest. We define rearrest as experiencing a new arrest before adjudication of the focal arrest, with detained defendants being assigned zero values for the purposes of this table. Arrest charge categories reflect the most serious criminal charge for which a person was arrested, using the FBI Uniform Crime Reporting hierarchy rule in cases where someone is arrested and charged with multiple offenses. For analyses of variance for the test of the joint null hypothesis that the difference in means across each variable is zero, see Online Appendix Table A.II.

overrepresents men (78.7%) and Black residents (69.4%). The average age of arrestees is 31.8 years. Judges detain 23.3% of cases, and in 25.1% of arrests the person is rearrested before their case is resolved (about one-third of those released). Randomization of arrestees to the training versus validation data sets seems to have been successful, as shown in Table I. None of the pairwise comparisons has a p-value below .05 (see Online Appendix Table A.II). A permutation multivariate analysis of variance test of the joint null hypothesis that the training-validation differences for all variables are all zero yields $p = .963$.[32] A test for the same joint null hypothesis for the differences between the training sample and the lock-box hold-out data set (out of sample by individual) yields a test statistic of $p = .537$.

III.C. Human Labels

The administrative data capture many key features of each case but omit some other important ones. We solve these data insufficiency problems through a series of human intelligence tasks (HITs), which involve having study subjects on one of two possible platforms (Amazon's Mechanical Turk or Prolific) assign labels to each case from looking at the mug shots. More details are in Online Appendix Table A.III. We use data from these HITs mostly to understand how the algorithm's predictions relate to already-known determinants of human decision making, and hence the degree to which the algorithm is discovering something novel.

One set of HITs filled in demographic-related data: ethnicity; skin tone (since people are often stereotyped on skin color, or "colorism"; Hunter 2007), reported on an 18-point scale; the degree to which defendants appear more stereotypically Black on a 9-point scale (Eberhardt et al. 2006 show this affects criminal justice decisions); and age, to compare to administrative data for label quality checks.[33] Because demographics tend to be easy

[32] Using the data on arrests up to July 17, 2019, we randomly reassign arrestees to three groups of similar size to our training, validation, and lock-box hold-out data sets, convert the data to long format (with one row for each arrest-and-variable) and calculate an F-test statistic for the joint null hypothesis that the difference in baseline characteristics are all zero, clustering standard errors by arrestee. We store that F-test statistic, rerun this procedure 1,000 times, and then report the share of splits with an F-statistic larger than the one observed for the original data partition.

[33] For an example HIT task, see Online Appendix Figure A.II.

for people to see in images, we collect just one label per image for each of these variables. To confirm one label is enough, we repeated the labeling task for 100 images but collected 10 labels for each image; we see that additional labels add little information.[34] Another data quality check comes from the fact that the distributions of skin color ratings do systematically differ by defendant race (Online Appendix Figure A.III).

A second type of HIT measured facial features that previous psychology research has shown affect human judgments. The specific set of facial features we focus on come from the influential study by Oosterhof and Todorov (2008) of people's perceptions of the facial features of others. When subjects are asked to provide descriptions of different faces, principal components analysis suggests just two dimensions account for about 80% of the variation: (i) trustworthiness and (ii) dominance. We also collected data on two other facial features shown to be associated with real-world decisions like hiring or whom to vote for: (iii) attractiveness and (iv) competence (Frieze, Olson, and Russell 1991; Little, Jones, and DeBruine 2011; Todorov and Oh 2021).[35]

We asked subjects to rate images for each of these psychological features on a nine-point scale. Because psychological features may be less obvious than demographic features, we collected three labels per training–data set image and five per validation–data set image.[36] There is substantial variation in the ratings that subjects assign to different images for each feature (see Online Appendix Figure A.VI). The ratings from different subjects for the same feature and image are highly correlated: interrater reliability measures (Cronbach's α) range from 0.87 to 0.98 (Online Appendix Figure A.VII), similar to those reported in

[34] For age and skin tone, we calculated the average pairwise correlation between two labels sampled (without replacement) from the 10 possibilities, repeated across different random pairs. The Pearson correlation was 0.765 for skin tone, 0.741 for age, and between age assigned labels versus administrative data, 0.789. The maximum correlation between the average of the first k labels collected and the $k + 1$ label is not all that much higher for $k = 1$ than $k = 9$ (0.733 versus 0.837).

[35] For an example of the consent form and instructions given to labelers, see Online Appendix Figures A.IV and A.V.

[36] We actually collected at least three and at least five, but the averages turned out to be very close to the minimums, equal to 3.17 and 5.07, respectively.

studies like Oosterhof and Todorov (2008).[37] The information gain from collecting more than a few labels per image is modest.[38] For summary statistics, see Online Appendix Table A.IV.

Finally, we also tried to capture people's implicit or tacit understanding of the determinants of judges' decisions by asking subjects to predict which mug shot out of a pair would be detained, with images in each pair matched on gender, race, and five-year age brackets.[39] We incentivized study subjects for correct predictions and gave them feedback over the course of the 50 image pairs to facilitate learning. We treat the first 10 responses per subject as a "learning set" that we exclude from our analysis.

IV. The Surprising Importance of the Face

The first step of our hypothesis generation procedure is to build an algorithmic model of some behavior, which in our case is the judge's detention decision. A sizable share of the predictable variation in judge decisions comes from a surprising source: the defendant's face. Facial features implicated by past research explain just a modest share of this predictable variation. The algorithm seems to have found a novel discovery.

IV.A. What Drives Judge Decisions?

We begin by predicting judge pretrial detention decisions ($y = 1$ if detain, $y = 0$ if release) using all the inputs available (x). We use the training data set to construct two separate models for the two types of data available. We apply gradient-boosted decision trees to predict judge decisions using the structured administrative data (current charge, prior record, age, gender), $m_s(x)$; for the unstructured data (raw pixel values from the mug shots), we train a convolutional neural network, $m_u(x)$. Each model returns an estimate of y (a predicted detention probability) for a given x. Because these initial steps of our procedure use

[37] For example, in Oosterhof and Todorov (2008), Supplemental Materials Table S2, they report Cronbach's α values of 0.95 for attractiveness, and 0.93 for both trustworthy and dominant.

[38] See Online Appendix Figure A.VIII, which shows that the change in the correlation between the $(k + 1)$th label with the mean of the first k labels declines after three labels.

[39] For an example, see Online Appendix Figure A.IX.

standard machine learning methods, we relegate their discussion to the Online Appendix.

We pool the signal from both models to form a single weighted-average model $m_p(x) = [\hat{\beta}_s m_s(x) + \hat{\beta}_u m_u(x)]$ using a so-called stacking procedure where the data are used to estimate the relevant weights.[40] Combining structured and unstructured data is an active area of deep-learning research, often called fusion modeling (Yuhas, Goldstein, and Sejnowski 1989; Lahat, Adali, and Jutten 2015; Ramachandram and Taylor 2017; Baltrušaitis, Ahuja, and Morency 2019). We have tried several of the latest fusion architectures; none improve on our ensemble approach.

Judge decisions do have some predictable structure. We report predictive performance as the area under the receiver operating characteristic curve, or AUC, which is a measure of how well the algorithm rank-orders cases with values from 0.5 (random guessing) to 1.0 (perfect prediction). Intuitively, AUC can be thought of as the chance that a uniformly randomly selected detained defendant has a higher predicted detention likelihood than a uniformly randomly selected released defendant. The algorithm built using all candidate features, $m_p(x)$, has an AUC of 0.780 (see Online Appendix Figure A.X).

What is the algorithm using to make its predictions? A single type of input captures a sizable share of the total signal: the defendant's face. The algorithm built using only the mug shot image, $m_u(x)$, has an AUC of 0.625 (see Online Appendix Figure A.X). Since an AUC of 0.5 represents random prediction, in AUC terms the mug shot accounts for $\frac{0.625-0.5}{0.780-0.5} = 44.6\%$ of the predictive signal about judicial decisions.

Another common way to think about predictive accuracy is in R^2 terms. While our data are high dimensional (because the facial image is a high-dimensional object), the algorithm's prediction of the judge's decision based on the facial image, $m_u(x)$, is a scalar and can be easily included in a familiar regression framework. Like AUC, measures like R^2 and mean squared error capture how well a model rank-orders observations by predicted probabilities,

[40] We use the validation data set to estimate $\hat{\beta}$ and then evaluate the accuracy of $m_p(x)$. Although this could lead to overfitting in principle, since we are only estimating a single parameter, this does not matter much in practice; we get very similar results if we randomly partition the validation data set by arrestee, use a random 30% of the validation data set to estimate the weights, then measure predictive performance in the other random 70% of the validation data set.

but R^2, unlike AUC, also captures how close predictions are to observed outcomes (calibration).[41] The R^2 from regressing y against $m_s(x)$ and $m_u(x)$ in the validation data is 0.11. Regressing y against $m_u(x)$ alone yields an R^2 of 0.03. So depending on how we measure predictive accuracy, around a quarter ($\frac{0.03}{0.11} = 27.3\%$) to a half (44.6%) of the predicted signal about judges' decisions is captured by the face.

Average differences are another way to see what drives judges' decisions. For any given feature x_k, we can calculate the average detention rate for different values of the feature. For example, for the variable measuring whether the defendant is male ($x_k = 1$) versus female ($x_k = 0$), we can calculate and plot $E[y \,|\, x_k = 1]$ versus $E[y \,|\, x_k = 0]$. As shown in Online Appendix Figure A.XI, the difference in detention rates equals 4.8 percentage points for those arrested for violent versus nonviolent crimes, 10.2 percentage points for men versus women, and 4.3 percentage points for bottom versus top quartile of skin tone, which are all sizable relative to the baseline detention rate of 23.3% in our validation data set. By way of comparison, average detention rates for the bottom versus top quartile of the mug shot algorithm's predictions, $m_u(x)$, differ by 20.4 percentage points.

In what follows, we seek to understand more about the mug shot–based prediction of the judge's decision, which we refer to simply as $m(x)$ in the remainder of the article.

IV.B. *Judicial Error?*

So far we have shown that the face predicts judges' behavior. Are judges right to use face information? To be precise, by "right" we do not mean a broader ethical judgment; for many reasons, one could argue it is never ethical to use the face. But suppose we take a rather narrow (exceedingly narrow) formulation of "right." Recall the judge is meant to make jailing decisions based on the defendant's risk. Is the use of these facial characteristics consistent with that objective? Put differently, if we account for defendant risk differences, do these facial characteristics still predict judge decisions? The fact that judges rely on the face in making detention decisions is in itself a striking insight regardless of whether

[41] The mean squared area for a linear probability model's predictions is related to the Brier score (Brier 1950). For a discussion of how this relates to AUC and calibration, see Murphy (1973).

the judges use appearance as a proxy for risk or are committing a cognitive error.

At first glance, the most straightforward way to answer this question would be to regress rearrest against the algorithm's mug shot–based detention prediction. That yields a statistically significant relationship: The coefficient (and standard error) for the mug shot equals 0.6127 (0.0460) with no other explanatory variables in the regression versus 0.5735 (0.0521) with all the explanatory variables (as in the final column, Table III). But the interpretation here is not so straightforward.

The challenge of interpretation comes from the fact that we have only measured crime rates for the released defendants. The problem with having measured crime, not actual crime, is that whether someone is charged with a crime is itself a human choice, made by police. If the choices police make about when to make an arrest are affected by the same biases that might afflict judges, then measured rearrest rates may correlate with facial characteristics simply due to measurement bias. The problem created by having measures of rearrest only for released defendants is that if judges have access to private information (defendant characteristics not captured by our data set), and judges use that information to inform detention decisions, then the released and detained defendants may be different in unobservable ways that are relevant for rearrest risk (Kleinberg et al. 2018).

With these caveats in mind, at least we can perform a bounding exercise. We created a predictor of rearrest risk (see Online Appendix B) and then regress judges' decisions on predicted rearrest risk. We find that a one-unit change in predicted rearrest risk changes judge detention rates by 0.6103 (standard error 0.0213). By comparison, we found that a one-unit change in the mug shot (by which we mean the algorithm's mug shot–based prediction of the judge detention decision) changes judge detention rates by 0.6963 (standard error 0.0383; see Table III, column (1)). That means if the judges were reacting to the defendant's face only because the face is a proxy for rearrest risk, the difference in rearrest risk for those with a one-unit difference in the mug shot would need to be $\frac{0.6963}{0.6103} = 1.141$. But when we directly regress rearrest against the algorithm's mug shot–based detention prediction, we get a coefficient of 0.6127 (standard error 0.0460). Clearly $0.6127 < 1.141$; that is, the mug shot does not seem to be strongly related enough to rearrest risk to explain the judge's use of it in making detention decisions.[42]

Of course this leaves us with the second problem with our data: we only have crime data on the released. It is possible the relationship between the mug shot and risk could be very different among the 23.3% of defendants who are detained (which we cannot observe). Put differently, the mug shot–risk relationship among the 76.7% of the defendants who are released is 0.6127; and let A be the (unknown) mug shot–risk relationship among the jailed. What we really want to know is the mug shot–risk relationship among all defendants, which equals $(0.767 \cdot 0.6127) + (0.233 \cdot A)$. For this mug shot–risk relationship among all defendants to equal 1.141, A would need to be 2.880, nearly five times as great among the detained defendants as among the released. This would imply an implausibly large effect of the mug shot on rearrest risk relative to the size of the effects on rearrest risk of other defendant characteristics.[43]

In addition, the results from Section VI.B call into question that these characteristics are well-understood proxies for risk. As we show there, experts who understand pretrial (public defenders and legal aid society staff) do not recognize the signal about judge decision making that the algorithm has discovered in the mug shot. These considerations as a whole—that measured rearrest is itself biased, the bounding exercise, and the failure of experts to recreate this signal—together lead us to tentatively conclude that it is unlikely that what the algorithm is finding in the face is merely a well-understood proxy for risk, but reflects errors in the judicial decision-making process. Of course, that presumption is not essential for the rest of the article, which asks: what exactly has the algorithm discovered in the face?

[42] Note how this comparison helps mitigate the problem that police arrest decisions could depend on a person's face. When we regress rearrest against the mug shot, that estimated coefficient may be heavily influenced by how police arrest decisions respond to the defendant's appearance. In contrast when we regress judge detention decisions against predicted rearrest risk, some of the variation across defendants in rearrest risk might come from the effect of the defendant's appearance on the probability a police officer makes an arrest, but a great deal of the variation in predicted risk presumably comes from people's behavior.

[43] The average mug shot–predicted detention risk for the bottom and top quartiles equal 0.127 and 0.332; that difference times 2.880 implies a rearrest risk difference of 59.0 percentage points. By way of comparison, the difference in rearrest risk between those who are arrested for a felony crime rather than a less serious misdemeanor crime is equal to just 7.8 percentage points.

IV.C. Is the Algorithm Discovering Something New?

Previous studies already tell us a number of things about what shapes the decisions of judges and other people. For example, we know people stereotype by gender (Avitzour et al. 2020), age (Neumark, Burn, and Button 2016; Dahl and Knepper 2020), and race or ethnicity (Bertrand and Mullainathan 2004; Arnold, Dobbie, and Yang 2018; Arnold, Dobbie, and Hull 2020; Fryer 2020; Hoekstra and Sloan 2022; Goncalves and Mello 2021). Is the algorithm just rediscovering known determinants of people's decisions, or discovering something new? We address this in two ways. We first ask how much of the algorithm's predictions can be explained by already-known features (Table II). We then ask how much of the algorithm's predictive power in explaining actual judges' decisions is diminished when we control for known factors (Table III). We carry out both analyses for three sets of known facial features: (i) demographic characteristics, (ii) psychological features, and (iii) incentivized human guesses.[44]

Table II, columns (1)–(3) show the relationship of the algorithm's predictions to demographics. The predictions vary enormously by gender (men have predicted detention likelihoods 11.9 percentage points higher than women), less so by age,[45] and by different indicators of race or ethnicity. With skin tone scored on a 0−1 continuum, defendants whom independent raters judge to be at the lightest end of the continuum are 4.4 percentage points less likely to be detained than those rated to have the darkest skin tone (column (3)). Conditional on skin tone, Black defendants have a 1.9 percentage point lower predicted likelihood of detention compared with whites.[46]

[44] In our main exhibits, we impose a simple linear relationship between the algorithm's predicted detention risk and known facial features like age or psychological variables, for ease of presentation. We show our results are qualitatively similar with less parametric specifications in Online Appendix Tables A.VI, A.VII, and A.VIII.

[45] With a coefficient value of 0.0006 on age (measured in years), the algorithm tells us that even a full decade's difference in age has 5% the impact on detention likelihood compared to the effects of gender ($10 \times 0.0006 = 0.6$ percentage point higher likelihood of detention, versus 11.9 percentage points).

[46] Online Appendix Table A.V shows that Hispanic ethnicity, which we measure from subject ratings from looking at mug shots, is not statistically significantly related to the algorithm's predictions. Table II, column (2) showed that conditional on gender, Black defendants have slightly higher predicted detention odds than white defendants (0.3 percentage points), but this is not quite

TABLE II

IS THE ALGORITHM REDISCOVERING KNOWN FACIAL FEATURES?

	Dependent variable Algorithmic judge detain prediction				
	(1)	(2)	(3)	(4)	(5)
Male	0.1186***	0.1179***	0.1153***	0.1138***	0.1140***
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)
Age		0.0006***	0.0006***	0.0003***	0.0003***
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		0.0029	−0.0185***	−0.0168***	−0.0171***
		(0.0023)	(0.0037)	(0.0036)	(0.0036)
Asian		−0.0204*	−0.0232**	−0.0210*	−0.0216*
		(0.0115)	(0.0115)	(0.0114)	(0.0114)
Indigenous American		0.0103	0.0061	0.0135	0.0126
		(0.0241)	(0.0240)	(0.0238)	(0.0238)
Skin tone			−0.0441***	−0.0411***	−0.0417***
			(0.0059)	(0.0058)	(0.0058)
Attractiveness				−0.0055***	−0.0051***
				(0.0016)	(0.0016)
Competence				−0.0091***	−0.0087***
				(0.0017)	(0.0017)
Dominance				0.0037***	0.0030**
				(0.0012)	(0.0012)
Trustworthiness				−0.0048***	−0.0041**
				(0.0016)	(0.0016)
Human guess					0.0399***
					(0.0062)
Constant	0.1595***	0.1391***	0.1771***	0.2393***	0.2173***
	(0.0022)	(0.0039)	(0.0064)	(0.0089)	(0.0095)
Observations	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.1954	0.1992	0.2038	0.2195	0.2228

Notes. The table presents the results of regressing an algorithmic prediction of judge detention decisions against each of the different explanatory variables as listed in the rows, where each column represents a different regression specification (the specific explanatory variables in each regression are indicated by the filled-in coefficients and standard errors in the table). The algorithm was trained using mug shots from the training data set; the regressions reported here are carried out using data from the validation data set. Data on skin tone, attractiveness, competence, dominance, and trustworthiness comes from asking subjects to assign feature ratings to mug shot images from the Mecklenburg County, NC, Sheriff's Office public website (see the text). The human guess about the judges' decision comes from showing workers on the Prolific platform pairs of mug shot images and asking them to report which defendant they believe the judge would be more likely to detain. Regressions follow a linear probability model and also include indicators for unknown race and unknown gender. * $p < .1$; ** $p < .05$; *** $p < .01$.

TABLE III

DOES THE ALGORITHM PREDICT JUDGE BEHAVIOR AFTER CONTROLLING FOR KNOWN FACTORS?

	Dependent variable: Judge detain decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algo judge detain prediction	0.6963***					0.6262***	0.6171***
	(0.0383)					(0.0433)	(0.0434)
Male		0.1040***	0.0978***		0.0940***	0.0228*	0.0244**
		(0.0105)	(0.0106)		(0.0108)	(0.0117)	(0.0117)
Age		−0.0008**	−0.0009**		−0.0013***	−0.0015***	−0.0015***
		(0.0004)	(0.0004)		(0.0004)	(0.0004)	(0.0004)
Black		−0.0139	−0.0651***		−0.0618***	−0.0513***	−0.0521***
		(0.0098)	(0.0156)		(0.0156)	(0.0154)	(0.0154)
Asian		−0.0753	−0.0818*		−0.0754	−0.0623	−0.0638
		(0.0490)	(0.0490)		(0.0489)	(0.0484)	(0.0484)
Indigenous American		0.0626	0.0524		0.0670	0.0585	0.0568
		(0.1024)	(0.1023)		(0.1021)	(0.1011)	(0.1010)
Skin tone			−0.1059***		−0.1004***	−0.0747***	−0.0762***
			(0.0251)		(0.0251)	(0.0249)	(0.0249)
Attractiveness				−0.0017	−0.0053	−0.0019	−0.0011
				(0.0063)	(0.0067)	(0.0067)	(0.0067)
Competence				−0.0192***	−0.0207***	−0.0150**	−0.0144**
				(0.0073)	(0.0072)	(0.0072)	(0.0072)
Dominance				0.0160***	0.0095*	0.0071	0.0057
				(0.0050)	(0.0051)	(0.0051)	(0.0051)

TABLE III

CONTINUED

	Dependent variable: Judge detain decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Trustworthiness				-0.0190^{***}	-0.0135^{*}	−0.0105	−0.0092
				(0.0070)	(0.0071)	(0.0070)	(0.0070)
Human guess							0.0852^{***}
							(0.0265)
Constant	0.0576^{***}	0.1868^{***}	0.2780^{***}	0.3054^{***}	0.3928^{***}	0.2429^{***}	0.1981^{***}
	(0.0106)	(0.0165)	(0.0272)	(0.0258)	(0.0381)	(0.0391)	(0.0415)
Naive-AUC	0.625	0.56	0.571	0.549	0.586	0.633	0.635
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.0331	0.0101	0.0119	0.0049	0.0162	0.0370	0.0380

Notes. This table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables in the validation set described in Table I. Each row represents a different explanatory variable for the regression, while each column reports the results of a separate regression with different combinations of explanatory variables (as indicated by the filled-in coefficients and standard errors in the table). The algorithmic predictions of the judges' detain decision come from our convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. Measures of defendant demographics and current arrest charge come from government administrative data obtained from a combination of Mecklenburg County, NC, and state agencies. Measures of skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mug shot images (see the text). Human guess variable comes from showing subjects pairs of mug shot 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 and unknown gender. * $p < .1$; ** $p < .05$; *** $p < .01$.

Table II, column (4) shows how the algorithm's predictions relate to facial features implicated by past psychological studies as shaping people's judgments of one another. These features also help explain the algorithm's predictions of judges' detention decisions: people judged by independent raters to be one standard deviation more attractive, competent, or trustworthy have lower predicted likelihood of detention equal to 0.55, 0.91, and 0.48 percentage points, respectively, or 2.2%, 3.6%, and 1.8% of the base rate.[47] Those whom subjects judge are one standard deviation more dominant-looking have a higher predicted likelihood of detention of 0.37 percentage points (or 1.5%).

How do we know we have controlled for everything relevant from past research? The literature on what shapes human judgments in general is vast; perhaps there are things that are relevant for judges' decisions specifically that we have inadvertently excluded? One way to solve this problem would be to do a comprehensive scan of past studies of human judgment and decision making, and then decide which results from different non–criminal justice contexts might be relevant for criminal justice. But that itself is a form of human-driven hypothesis generation, bringing us right back to where we started.

To get out of this box, we take a different approach. Instead of enumerating individual characteristics, we ask people to embody their beliefs in a guess, which ought to be the compound of all these characteristics. Then we can ask whether the algorithm has rediscovered this human guess (and later whether it has discovered more). We ask independent subjects to look at pairs of mug shots matched by gender, race, and five-year age bins and forecast which defendant is more likely to be detained by a judge. We provide a financial incentive for accurate guesses to increase the

significant (t = 1.3). Online Appendix Table A.V, column (1) shows that conditioning on Hispanic ethnicity and having stereotypically Black facial features—as measured in Eberhardt et al. (2006)—increases the size of the Black-white difference in predicted detention odds (now equal to 0.8 percentage points) as well as the difference's statistical significance ($t = 2.2$).

[47] This comes from multiplying the effect of each 1 unit change in our 9-point scale associated, equal to 0.55, 0.91, and 0.48 percentage points, respectively, with the standard deviation of the average label for each psychological feature for each image, which equal 0.923, 0.911, and 0.844, respectively.

chances that subjects take the exercise seriously.[48] We also provide subjects with an opportunity to learn by showing subjects 50 image pairs with feedback after each pair about which defendant the judge detained. We treat the first 10 image pairs from each subject as learning trials and only use data from the last 40 image pairs. This approach is intended to capture anything that influences judges' decisions that subjects could recognize, from subtle signs of things like socioeconomic status or drug use or mood, to things people can recognize but not articulate.

It turns out subjects are modestly good at this task (Table II). Participants guess which mug shot is more likely to be detained at a rate of 51.4%, which is different to a statistically significant degree from the 50% random-guessing threshold. When we regress the algorithm's predicted detention rate against these subject guesses, the coefficient is 3.99 percentage points, equal to 17.1% of the base rate.

The findings in Table II are somewhat remarkable. The only input the algorithm had access to was the raw pixel values of each mug shot, yet it has rediscovered findings from decades of previous research and human intuition.

Interestingly, these features collectively explain only a fraction of the variation in the algorithm's predictions: the R^2 is only 0.2228. That by itself does not necessarily mean the algorithm has discovered additional useful signal. It is possible that the remaining variation is prediction error—components of the prediction that do not explain actual judges' decisions.

In Table III, we test whether the algorithm uncovers any additional signal for actual judge decisions, above and beyond the influence of these known factors. The algorithm by itself produces an R^2 of 0.0331 (column (1)), substantially higher than all previously known features taken together, which produce an R^2 of 0.0162 (column (5)), or the human guesses alone which produce an R^2 of 0.0025 (so we can see the algorithm is much better at predicting detention from faces than people are). Another way to see that the algorithm has detected signal above and beyond these known features is that the coefficient on the algorithm prediction when included alone in the regression, 0.6963 (column (1)),

[48] As discussed in Online Appendix Table A.III, we offer subjects a \$3.00 base rate for participation plus an incentive of 5 cents per correct guess. With 50 image pairs shown to each participant, they could increase their earnings by another \$2.50, or up to 83% above the base compensation.

changes only modestly when we condition on everything else, now equal to 0.6171 (column (7)). The algorithm seems to have discovered some novel source of signal that better predicts judge detention decisions.[49]

V. Algorithm-Human Communication

The algorithm has made a discovery: something about the defendant's face explains judge decisions, above and beyond the facial features implicated by existing research. But what is it about the face that matters? Without an answer, we are left with a discovery of an unsatisfying sort. We have simply replaced one black box hypothesis generation procedure (human creativity) with another (the algorithm). In what follows we demonstrate how existing methods like saliency maps cannot solve this challenge in our application and then discuss our solution to that problem.

V.A. *The Challenge of Explanation*

The problem of algorithm-human communication stems from the fact that we cannot simply look inside the algorithm's "black box" and see what it is doing because $m(x)$, the algorithmic predictor, is so complicated. A common solution in computer science is to forget about looking inside the algorithmic black box and focus instead on drawing inferences from curated outputs of that box. Many of these methods involve gradients: given a prediction function $m(x)$, we can calculate the gradient $\nabla m(x) = \frac{\mathrm{d}m}{\mathrm{d}x}(x)$. This lets us determine, at any input value, what change in the input vector maximally changes the prediction.[50] The idea of gradients is useful for image classification tasks because it allows us to tell

[49] Table III gives us another way to see how much of previously known features are rediscovered by the algorithm. That the algorithm's prediction plus all previously known features yields an R^2 of just 0.0380 (column (7)), not much larger than with the algorithm alone, suggests the algorithm has discovered most of the signal in these known features. But not necessarily all: these other known features often do remain statistically significant predictors of judges' decisions even after controlling for the algorithm's predictions (last column). One possible reason is that, given finite samples, the algorithm has only imperfectly reconstructed factors such as "age" or "human guess." Controlling for these factors directly adds additional signal.

[50] Imagine a linear prediction function like $m(x_1, x_2) = \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$. If our best estimates suggested $\widehat{\beta}_2 = 0$, the maximum change to the prediction comes from incrementally changing x_1.

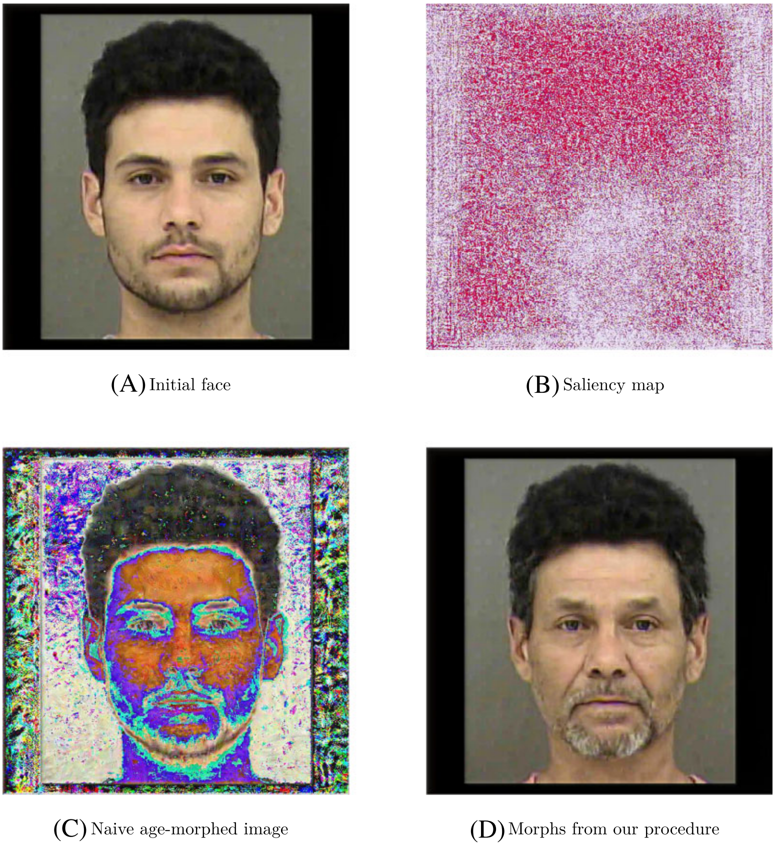
which pixel image values are most important for changing the predicted outcome.

For example, a widely used method known as saliency maps uses gradient information to highlight the specific pixels that are most important for predicting the outcome of interest (Baehrens et al. 2010; Simonyan, Vedaldi, and Zisserman 2014). This approach works well for many applications like determining whether a given picture contains a given type of animal, a common task in ecology (Norouzzadeh et al. 2018). What distinguishes a cat from a dog? A saliency map for a cat detector might highlight pixels around, say, the cat's head: what is most cat-like is not the tail, paws, or torso, but the eyes, ears, and whiskers. But more complicated outcomes of the sort social scientists study may depend on complicated functions of the entire image.

Even if saliency maps were more selective in highlighting pixels in applications like ours, for hypothesis generation they also suffer from a second limitation: they do not convey enough information to enable people to articulate interpretable hypotheses. In the cat detector example, a saliency map can tell us that something about the cat's (say) whiskers are key for distinguishing cats from dogs. But what about that feature matters? Would a cat look more like a dog if its whiskers were longer? Or shorter? More (or less?) even in length? People need to know not just what features matter but how they must change to change the prediction. For hypothesis generation, the saliency map undercommunicates with humans.

To test the ability of saliency maps to help with our application, we focused on a facial feature that people already understand and can easily recognize from a photo: age. We first build an algorithm that predicts each defendant's age from their mug shot. For a representative image, as in the top left of Figure III, we can highlight which pixels are most important for predicting age, shown in the top right.[51] A key limitation of saliency maps is easy to see: because age (like many human facial features) is a function of almost every part of a person's face, the saliency map highlights almost everything.

[51] As noted already, to avoid contributing to the stereotyping of minorities in discussions of crime, in our exhibits we show images for non-Hispanic white men, although in our HITs we use images representative of the larger defendant population.



(A) Initial face

(B) Saliency map

(C) Naive age-morphed image

(D) Morphs from our procedure

FIGURE III

Candidate Algorithm-Human Communication Vehicles for a Known Facial Feature: Age

Panel A shows a randomly selected point in the GAN latent space for a non-Hispanic white male defendant. Panel B shows a saliency map that highlights the pixels that are most important for an algorithmic model that predicts the defendant's age from the mug shot image. Panel C shows an image changed or "morphed" in the direction of older age, based on the gradient of the image-based age prediction, using the "naive" morphing procedure that does not constrain the new image to lie on the face manifold (see the text). Panel D shows the image morphed to the maximum age using our actual preferred morphing procedure.

An alternative to simply highlighting high-leverage pixels is to change them in the direction of the gradient of the predicted outcome, to—ideally—create a new face that now has a different predicted outcome, what we call "morphing." This new image

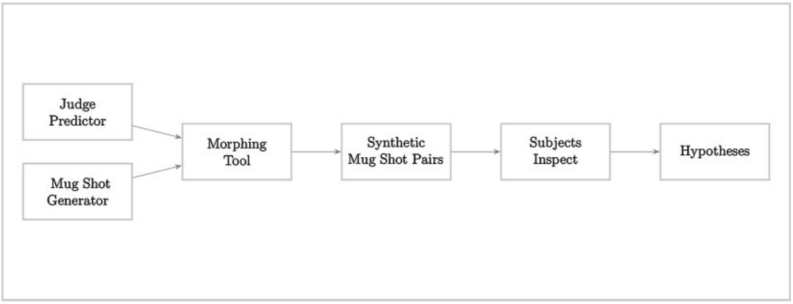


FIGURE IV

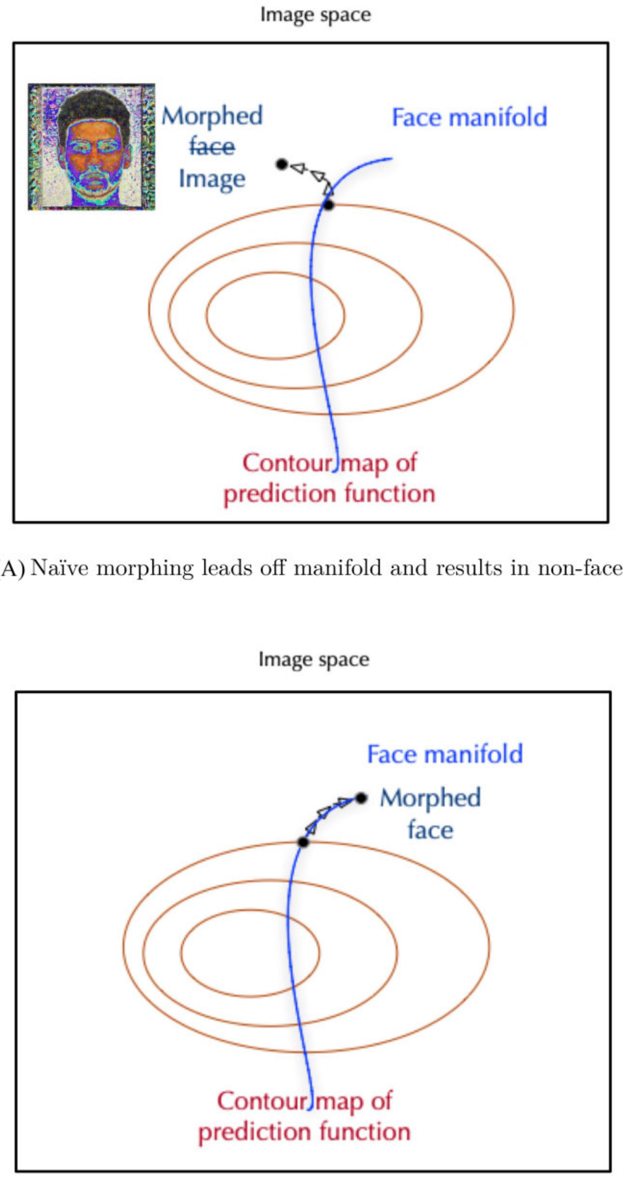
Hypothesis Generation Pipeline

The diagram illustrates all the algorithmic components in our procedure by presenting a full pipeline for algorithmic interpretation.

answers the counterfactual question: "How would this person's face change to increase their predicted outcome?" Our approach builds on the ability of people to comprehend ideas through comparisons, so we can show morphed image pairs to subjects to have them name the differences that they see. Figure IV summarizes our semiautomated hypothesis generation pipeline. (For more details see Online Appendix B.) The benefit of morphed images over actual mug shot images is to isolate the differences across faces that matter for the outcome of interest. By reducing noise, morphing also reduces the risk of spurious discoveries.

Figure V illustrates how this morphing procedure works in practice and highlights some of the technical challenges that arise. Let the box in the top panel represent the space of all possible images—all possible combinations of pixel values for, say, a 512 × 512 image. Within this space, we can apply our mug shot–based predictor of the known facial feature, age, to identify all images with the same predicted age, as shown by the contour map of the prediction function. Imagine picking some random initial mug shot image. We could follow the gradient to find an image with a higher predicted value of the outcome y.

The challenge is that most points in this image space are not actually face images. Simply following the gradient will usually take us off the data distribution of face images, as illustrated abstractly in the top panel of Figure V. What this means in practice is shown in the bottom left panel of Figure III: the result is an image that has a different predicted outcome (in the figure,



(A) Naïve morphing leads off manifold and results in non-faces

(B) Our procedure stays on manifold and morphs are faces

FIGURE V

Morphing Images for Detention Risk On and Off the Face Manifold

FIGURE V

(*Continued*) The figure shows the difference between an unconstrained (naive) morphing procedure and our preferred new morphing approach. In both panels, the background represents the image space (set of all possible pixel values) and the blue line (color version available online) represents the set of all pixel values that correspond to any face image (the face manifold). The orange lines show all images that have the same predicted outcome (isoquants in predicted outcome). The initial face (point on the outermost contour line) is a randomly selected face in GAN face space. From there we can naively follow the gradients of an algorithm that predicts some outcome of interest from face images. As shown in Panel A, this takes us off the face manifold and yields a nonface image. Alternatively, with a model of the face manifold, we can follow the gradient for the predicted outcome while ensuring that the new image is again a realistic instance as shown in Panel B.

illustrated for age) but no longer looks like a real instance—that is, no longer looks like a realistic face image. This "naive" morphing procedure will not work without some way to ensure the new point we wind up on in image space corresponds to a realistic face image.

V.B. *Building a Model of the Data Distribution*

To ensure morphing leads to realistic face images, we need a model of the data distribution $p(x)$—in our specific application, the set of images that are faces. We rely on an unsupervised learning approach to this problem.[52] Specifically, we use generative adversarial networks (GANs), originally introduced to generate realistic new images for a variety of tasks (see Goodfellow et al. 2014).[53]

A GAN is built by training two algorithms that "compete" with each another, 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 which is which. A good discriminator pressures the generator to produce images that are harder to distinguish from real; in turn, a good generator pressures the classifier to get better at discriminating real from synthetic images. Data on actual faces

[52] Modeling $p(x)$ through a supervised learning task would involve assembling a large set of images, having subjects label each image for whether they contain a realistic face, and then predicting those labels using the image pixels as inputs. But this supervised learning approach is costly because it requires extensive annotation of a large training data set.

[53] Kaji, Manresa, and Pouliot (2020) and Athey et al. (2021, 2022) are recent uses of GANs in economics.

are used to train the discriminator, which results in the generator being trained as it seeks to fool the discriminator. With machine learning, the performance of 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 data distribution of faces. (Additional discussion of GANs in general and how we construct our GAN specifically are in Online Appendix B.)

To build our GAN and evaluate its expressiveness we use standard training metrics, which turn out to compare favorably to what we see with other widely used GAN models on other data sets (see Online Appendix B.C for details). A more qualitative way to judge our GAN comes from visual inspection; some examples of synthetic face images are in Figure II. Most importantly, the GAN we build (as is true of GANs in general) is not generic. GANs are specific. They do not generate "faces" but instead seek to match the distribution of pixel combinations in the training data. For example, our GAN trained using mug shots would never generate generic Facebook profile photos or celebrity headshots.

Figure V illustrates how having a model such as the GAN lets morphing stay on the data distribution of faces and produce realistic images. We pick a random point in the space of faces (mug shots) and then use the algorithmic predictor of the outcome of interest $m(x)$ to identify nearby faces that are similar in all respects except those relevant for the outcome. Notice this procedure requires that faces closer to one another in GAN latent space should look relatively more similar to one another to a human in pixel space. Otherwise we might make a small movement along the gradient and wind up with a face that looks different in all sorts of other ways that are irrelevant to the outcome. That is, we need the GAN not just to model the support of the data but also to provide a meaningful distance metric.

When we produce these morphs, what can possibly change as we morph? In principle there is no limit. The changes need not be local: features such as skin color, which involves many pixels, could change. So could features such as attractiveness, where the pixels that need to change to make a face more attractive vary from face to face: the "same" change may make one face more attractive and another less so. Anything represented in the face could change, as could anything else in the image beyond the face that matters for the outcome (if, for example, localities varied in both detention rates and the type of background they have someone stand in front of for mug shots).

In practice, though, there is a limit. What can change depends on how rich and expressive the estimated GAN is. If the GAN fails to capture a certain kind of face or a dimension of the face, then we are unlikely to be able to morph on that dimension. The morphing procedure is only as complete as the GAN is expressive. Assuming the GAN expresses a feature, then if $m(x)$ truly depends on that feature, morphing will likely display it. Nor is there any guarantee that in any given application the classifier $m(x)$ will find novel signal for the outcome y, or that the GAN successfully learns the data distribution (Nalisnick et al. 2018), or that subjects can detect and articulate whatever signal the classifier algorithm has discovered. Determining the general conditions under which our procedure will work is something we leave to future research. Whether our procedure can work for the specific application of judge decisions is the question to which we turn next.[54]

V.C. Validating the Morphing Procedure

We return to our algorithmic prediction of a known facial feature—age—and see what morphing by age produces as a way to validate or test our procedure. When we follow the gradient of the predicted outcome (age), by constraining ourselves to stay on the GAN's latent space of faces we wind up with a new age-morphed face that does indeed look like a realistic face image, as shown in the bottom right of Figure III. We seem to have successfully developed a model of the data distribution and a way to move around on that surface to create realistic new instances.

[54] Some ethical issues are worth considering. One is bias. With human hypothesis generation there is the risk people "see" an association that impugns some group yet has no basis in fact. In contrast our procedure by construction only produces empirically plausible hypotheses. A different concern is the vulnerability of deep learning to adversarial examples: tiny, almost imperceptible changes in an image changing its classification for the outcome y, so that mug shots that look almost identical (that is, are very "similar" in some visual image metric) have dramatically different $m(x)$. This is a problem because tiny changes to an image don't change the nature of the object; see Szegedy et al. (2013) and Goodfellow, Shlens, and Szegedy (2014). In practice such instances are quite rare in nature, indeed, so rare they usually occur only if intentionally (maliciously) generated.

To figure out if algorithm-human communication occurs, we run these age-morphed image pairs through our experimental pipeline (Figure IV). Our procedure is only useful if it is replicable—that is, if it does not depend on the idiosyncratic insights of any particular person. For that reason, the people looking at these images and articulating what they see should not be us (the investigators) but a sample of external, independent study subjects. In our application, we use Prolific workers (see Online Appendix Table A.III). Reliability or replicability is indicated by the agreement in the subject responses: lots of subjects see and articulate the same thing in the morphed images.

We asked subjects to look at 50 age-morphed image pairs selected at random from a population of 100 pairs, and told them the images in each pair differ on some hidden dimension but did not tell them what that was.[55] We asked subjects to guess which image expresses that hidden feature more, gave them feedback about the right answer, treated the first 10 image pairs as learning examples, and calculated accuracy on the remaining 40 images. Subjects correctly selected the older image 97.8% of the time.

The final step was to ask subjects to name what differs in image pairs. Making sense of these responses requires some way to group them into semantic categories. Each subject comment could include several concepts (e.g., “wrinkles, gray hair, tired”). We standardized these verbal descriptions by removing punctuation, using only lowercase characters, and removing stop words. We gave three research assistants not otherwise involved in the project these responses and asked them to create their own categories that would capture all the responses (see Online Appendix Figure A.XIII). We also gave them an illustrative subject comment and highlighted the different “types” of categories (descriptive physical features, i.e., “thick eyebrows,” descriptive impression category, i.e., “energetic,” but also an illustration of a category of comment that is too vague to lend itself to

[55] Online Appendix Figure A.XII gives an example of this task and the instructions given to participating subjects to complete it. Each subject was tested on 50 image pairs selected at random from a population of 100 images. Subjects were told that for every pair, one image was higher in some unknown feature, but not given details as to what the feature might be. As in the exercise for predicting detention, feedback was given immediately after selecting an image, and a 5 cent bonus was paid for every correct answer.

useful measurement, i.e., “ears”). In our validation exercise 81.5% of subject reports fall into the semantic categories of either age or the closely related feature of hair color.[56]

V.D. Understanding the Judge Detention Predictor

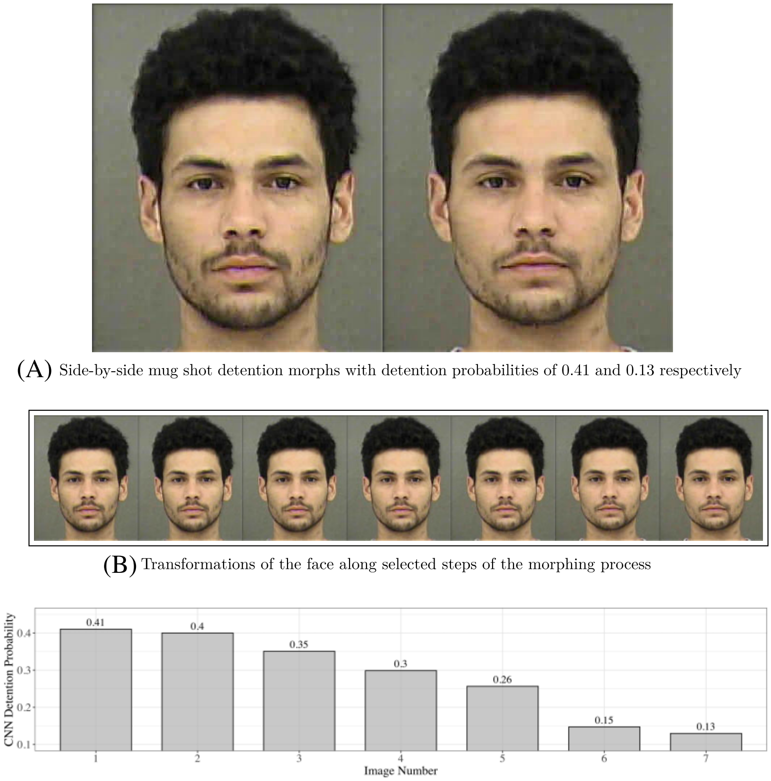
Having validated our algorithm-human communication procedure for the known facial feature of age, we are ready to apply it to generate a new hypothesis about what drives judge detention decisions. To do this we combine the mug shot algorithm predictor of judges’ detention decisions, $m(x)$, with our GAN of the data distribution of mug shot images, then create new synthetic image pairs morphed with respect to the likelihood the judge would detain the defendant (see Figure IV).

The top panel of Figure VI shows a pair of such images. Underneath we show an “image strip” of intermediate steps, along with each image’s predicted detention rate. With an overall detention rate of 23.3% in our validation data set, morphing takes us from about one-half the base rate (13%) up to nearly twice the base rate (41%). Additional examples of morphed image pairs are shown in Figure VII.

We showed 54 subjects 50 detention-risk-morphed image pairs each, asked them to predict which defendant would be detained, offered them financial incentives for correct answers,[57] and gave them feedback on the right answer. Online Appendix Figure A.XV shows how accurate subjects are as they get more practice across successive morphed image pairs. With the initial image-pair trials, subjects are not much better than random guessing, in the range of what we see when subjects look at pairs of actual mugshots (where accuracy is 51.4% across the final 40 mug shot pairs people see). But unlike what happens when subjects look at actual images, when looking at morphed image pairs subjects seem to quickly learn what the algorithm is trying to communicate to them. Accuracy increased by over 10 percentage points after 20 morphed image pairs and reached 67% after 30 image pairs. Compared to looking at actual mugshots, the morphing

[56] In principle this semantic grouping could be carried out in other ways, for example, with automated procedures involving natural-language processing.

[57] See Online Appendix Table A.III for a high-level description of this human intelligence task, and Online Appendix Figure A.XIV for a sample of the task and the subject instructions.



(A) Side-by-side mug shot detention morphs with detention probabilities of 0.41 and 0.13 respectively

(B) Transformations of the face along selected steps of the morphing process

(C) Detention probabilities for images in panel (b)

FIGURE VI

Illustration of Morphed Faces along the Detention Gradient

Panel A shows the result of selecting a random point on the GAN latent face space for a white non-Hispanic male defendant, then using our new morphing procedure to increase the predicted detention risk of the image to 0.41 (left) or reduce the predicted detention risk down to 0.13 (right). The overall average detention rate in the validation data set of actual mug shot images is 0.23 by comparison. Panel B shows the different intermediate images between these two end points, while Panel C shows the predicted detention risk for each of the images in the middle panel.

procedure accomplished its goal of making it easier for subjects to see what in the face matters most for detention risk.

We asked subjects to articulate the key differences they saw across morphed image pairs. The result seems to be a reliable hypothesis—a facial feature that a sizable share of subjects name. In the top panel of Figure VIII, we present a histogram

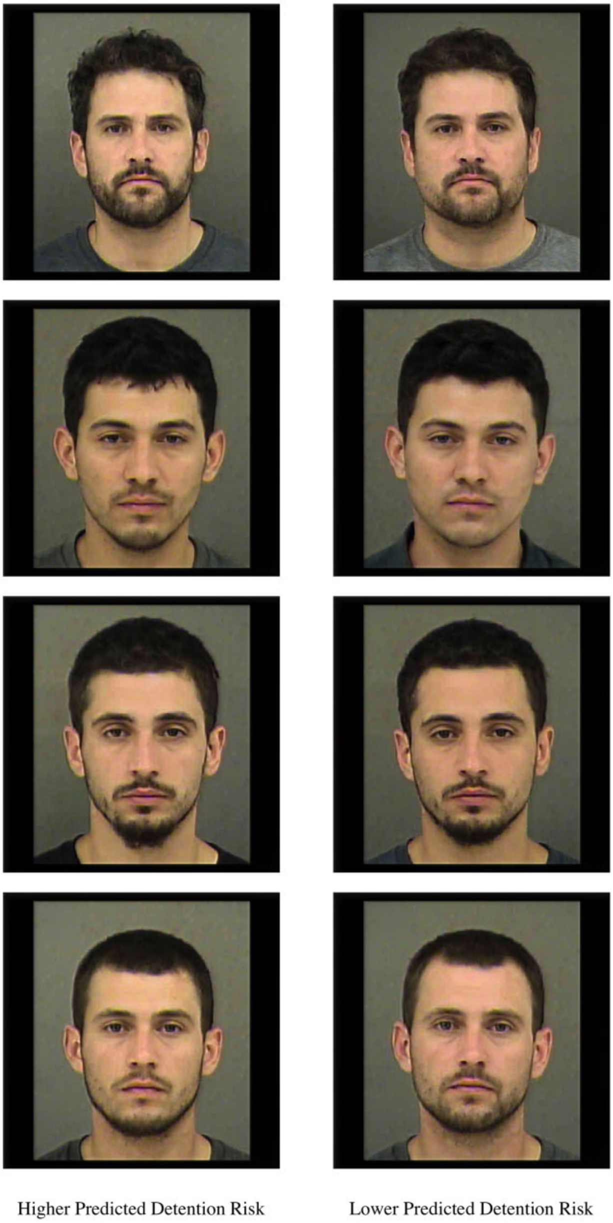
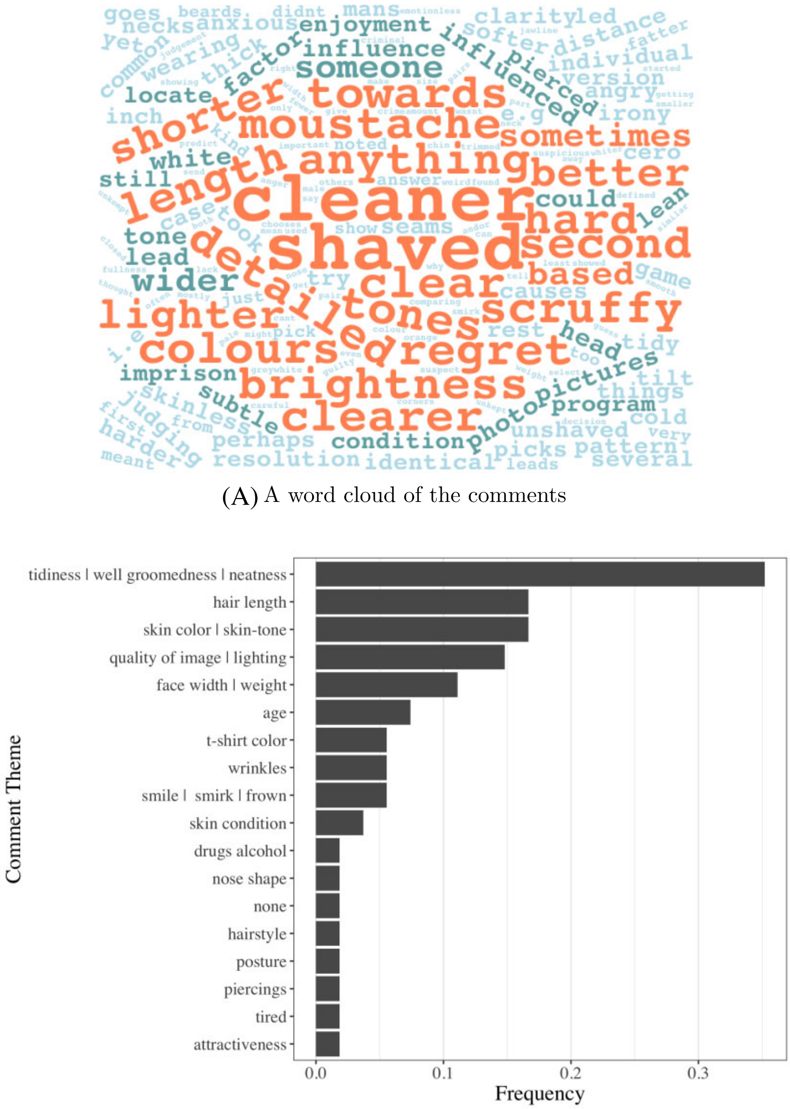

Higher Predicted Detention Risk
Lower Predicted Detention Risk

FIGURE VII

Examples of Morphing along the Gradients of the Face-Based Detention Predictor



(A) A word cloud of the comments

(B) Frequencies of comments by theme

FIGURE VIII

Subject Reports of What They See between Detention-Risk-Morphed Image Pairs

Panel A shows a word cloud of subject reports about what they see as the key difference between image pairs where one is a randomly selected point in the GAN latent space and the other is morphed in the direction of a higher predicted

FIGURE VIII

(*Continued*) detention risk. Words are approximately proportionately sized to the frequency of subject mentions. Panel B shows the frequency of semantic groupings of those open-ended subject reports (see the text for additional details).

of individual tokens (cleaned words from worker comments) in "word cloud" form, where word size is approximately proportional to frequency.[58] Some of the most common words are "shaved," "cleaner," "length," "shorter," "moustache," and "scruffy." To form semantic categories, we use a procedure similar to what we describe for our validation exercise for the known feature of age.[59] Grouping tokens into semantic categories, we see that nearly 40% of the subjects see and name a similar feature that they think helps explain judge detention decisions: how well-groomed the defendant is (see the bottom panel of Figure VIII).[60]

Can we confirm that what the subjects think the algorithm is seeing is what the algorithm actually sees? We asked a separate set of 343 independent subjects (MTurk workers) to label the 32,881 mug shots in our combined training and validation data sets for how well-groomed each image was perceived to be on a nine-point scale.[61] For data sets of our size, these labeling costs

[58] We drop every token of just one or two characters in length, as well as connector words without real meaning for this purpose, like "had," "the," and "and," as well as words that are relevant to our exercise but generic, like "jailed," "judge," and "image."

[59] We enlisted three research assistants blinded to the findings of this study and asked them to come up with semantic categories that captured all subject comments. Since each assistant mapped each subject comment to 5% of semantic categories on average, if the assistant mappings were totally uncorrelated, we would expect to see agreement of at least two assistant categorizations about 5% of the time. What we actually see is if one research assistant made an association, 60% of the time another assistant would make the same association. We assign a comment to a semantic category when at least two of the assistants agree on the categorization.

[60] Moreover what subjects see does not seem to be particularly sensitive to which images they see. (As a reminder, each subject sees 50 morphed image pairs randomly selected from a larger bank of 100 morphed image pairs). If we start with a subject who says they saw "well-groomed" in the morphed image pairs they saw, for other subjects who saw 21 or fewer images in common (so saw mostly different images) they also report seeing well-groomed 31% of the time, versus 35% among the population. We select the threshold of 21 images because this is the smallest threshold in which at least 50 pairs of raters are considered.

[61] See Online Appendix Table A.III and Online Appendix Figure A.XVI. This comes to a total of 192,280 individual labels, an average of 3.2 labels per image in the training set and an average of 10.8 labels per image in the validation set.

are fairly modest, but in principle those costs could be much more substantial (or even prohibitive) in some applications.

Table IV suggests algorithm-human communication has successfully occurred: our new hypothesis, call it $h_1(x)$, is correlated with the algorithm's prediction of the judge, $m(x)$. If subjects were mistaken in thinking they saw well-groomed differences across images, there would be no relationship between well-groomed and the detention predictions. Yet what we actually see is the R^2 from regressing the algorithm's predictions against well-groomed equals 0.0247, or 11% of the R^2 we get from a model with all the explanatory variables (0.2361). In a bivariate regression the coefficient (-0.0172) implies that a one standard deviation increase in well-groomed (1.0118 points on our 9-point scale) is associated with a decline in predicted detention risk of 1.74 percentage points, or 7.5% of the base rate. Another way to see the explanatory power of this hypothesis is to note that this coefficient hardly changes when we add all the other explanatory variables to the regression (equal to -0.0153 in the final column) despite the substantial increase in the model's R^2.

V.E. *Iteration*

Our procedure is iterable. The first novel feature we discovered, well-groomed, explains some—but only some—of the variation in the algorithm's predictions of the judge. We can iterate our procedure to generate hypotheses about the remaining residual variation as well. Note that the order in which features are discovered will depend on how important each feature is in explaining the judge's detention decision and on how salient each feature is to the subjects who are viewing the morphed image pairs. So explanatory power for the judge's decisions need not monotonically decline as we iterate and discover new features.

To isolate the algorithm's signal above and beyond what is explained by well-groomed, we wish to generate a new set of morphed image pairs that differ in predicted detention but hold well-groomed constant. That would help subjects see other novel features that might differ across the detention-risk-morphed images, without subjects getting distracted by differences in

Sampling labels from different workers on the same image, these ratings have a correlation of 0.14.

TABLE IV
CORRELATION BETWEEN WELL-GROOMED AND THE ALGORITHM'S PREDICTION

	Dependent variable: Algorithmic judge detain prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Well-groomed	−0.0172***	−0.0188***	−0.0184***	−0.0185***	−0.0158***	−0.0153***
	(0.0011)	(0.0010)	(0.0010)	(0.0010)	(0.0012)	(0.0012)
Male		0.1201***	0.1192***	0.1166***	0.1153***	0.1154***
		(0.0024)	(0.0024)	(0.0024)	(0.0025)	(0.0025)
Age			0.0003***	0.0002***	0.0002**	0.0002**
			(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black			0.0050**	−0.0168***	−0.0165***	−0.0168***
			(0.0023)	(0.0036)	(0.0036)	(0.0036)
Asian			−0.0138	−0.0165	−0.0153	−0.0160
			(0.0113)	(0.0113)	(0.0113)	(0.0113)
Indigenous American			0.0211	0.0169	0.0181	0.0172
			(0.0237)	(0.0236)	(0.0236)	(0.0236)
Skin tone				−0.0449***	−0.0437***	−0.0440***
				(0.0058)	(0.0058)	(0.0058)
Attractiveness					0.0006	0.0008
					(0.0016)	(0.0016)
Competence					−0.0062***	−0.0060***
					(0.0017)	(0.0017)
Dominance					0.0036***	0.0031**
					(0.0012)	(0.0012)

TABLE IV

CONTINUED

	Dependent variable: Algorithmic judge detain prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Trustworthiness					-0.0029^{*}	-0.0024
					(0.0016)	(0.0016)
Human guess						0.0339^{***}
						(0.0062)
Constant	0.3348^{***}	0.2486^{***}	0.2346^{***}	0.2736^{***}	0.2767^{***}	0.2568^{***}
	(0.0054)	(0.0051)	(0.0065)	(0.0082)	(0.0092)	(0.0099)
Observations	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.0247	0.2249	0.2262	0.2310	0.2337	0.2361

Notes. This table shows the results of estimating a linear probability specification regressing algorithmic predictions of judges' detain decision against different explanatory variables, using data from the validation set of cases from Mecklenburg County, NC. Each row of the table represents a different explanatory variable for the regression, while each column reports the results of a separate regression with different combinations of explanatory variables (as indicated by the filled-in coefficients and standard errors in the table). 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, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mug shot images (see the text). Human guess variable comes from showing subjects pairs of mug shot 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 and unknown gender. * $p < .1$; ** $p < .05$; *** $p < .01$.

well-groomed.[62] But iterating the procedure raises several technical challenges. To see these challenges, consider what would in principle seem to be the most straightforward way to orthogonalize, in the GAN's latent face space:

- use training data to build predictors of detention risk, $m(x)$, and the facial features to orthogonalize against, $h_1(x)$;
- pick a point on the GAN latent space of faces;
- collect the gradients with respect to $m(x)$ and $h_1(x)$;
- use the Gram-Schmidt process to move within the latent space toward higher predicted detention risk $m(x)$, but orthogonal to $h_1(x)$; and
- show new morphed image pairs to subjects, have them name a new feature.

The challenge with implementing this playbook in practice is that we do not have labels for well-groomed for the GAN-generated synthetic faces. Moreover, it would be infeasible to collect this feature for use in this type of orthogonalization procedure.[63] That means we cannot orthogonalize against well-groomed, only against predictions of well-groomed. And orthogonalizing with respect to a prediction is an error-prone process whenever the predictor is imperfect (as it is here).[64] The errors in the process accumulate as we take many morphing steps. Worse,

[62] It turns out that skin tone is another feature that is correlated with well-groomed, so we orthogonalize on that as well as well-groomed. To simplify the discussion, we use "well-groomed" as a stand-in for both features we orthogonalize against, well-groomed plus skin tone.

[63] To see why, consider the mechanics of the procedure. Since we orthogonalize as we create morphs, we would need labels at each morphing step. This would entail us producing candidate steps (new morphs), collecting data on each of the candidates, picking one that has the same well-groomed value, and then repeating. Moreover, until the labels are collected at a given step, the next step could not be taken. Since producing a final morph requires hundreds of such intermediate morphing steps, the whole process would be so time- and resource-consuming as to be infeasible.

[64] While we can predict demographic features like race and age (above/below median age) nearly perfectly, with AUC values close to 1, for predicting well-groomed, the mean absolute error of our OOS prediction is 0.63, which is plus or minus over half a slider value for this 9-point-scaled variable. One reason it is harder to predict well-groomed is because the labels, which come from human subjects looking at and labeling mug shots, are themselves noisy, which introduces irreducible error.

that accumulated error is not expected to be zero on average. Because we are morphing in the direction of predicted detention and we know predicted detention is correlated with well-groomed, the prediction error will itself be correlated with well-groomed.

Instead we use a different approach. We build a new detention-risk predictor with a curated training data set, limited to pairs of images matched on the features to be orthogonalized against. For each detained observation i (such that $y_i = 1$), we find a released observation j (such that $y_j = 0$) where $h_1(x_i) = h_1(x_j)$. In that training data set y is now orthogonal to $h_1(x)$, so we can use the gradient of the orthogonalized detention risk predictor to move in GAN latent space to create new morphed images with different detention odds but are similar with respect to well-groomed.[65] We call these "orthogonalized morphs," which we then feed into the experimental pipeline shown in Figure IV.[66] An open question for future work is how many iterations are possible before the dimensionality of the matching problem required for this procedure would create problems.

Examples from this orthogonalized image-morphing procedure are in Figure IX. Changes in facial features across morphed images are notably different from those in the first iteration of morphs as in Figure VI. From these examples, it appears possible that orthogonalization may be slightly imperfect; sometimes they show subtle differences in "well-groomed" and perhaps age. As with the first iteration of the morphing procedure, the second (orthogonalized) iteration of the procedure again generates images that vary substantially in their predicted risk, from 0.07 up to 0.27 (see Online Appendix Figure A.XVIII).

Still, there is a salient new signal: when presented to subjects they name a second facial feature, as shown in Figure X. We showed 52 subjects (Prolific workers) 50 orthogonalized morphed image pairs and asked them to name the differences they see. The word cloud shown in the top panel of Figure X shows that some of the most common terms reported by subjects include

[65] For additional details see Online Appendix Figure A.XVII and Online Appendix B.

[66] There are a few additional technical steps required, discussed in Online Appendix B. For details on the HIT we use to get subjects to name the new hypothesis from looking at orthogonalized morphs, and the follow-up HIT to generate independent labels for that new hypothesis or facial feature, see Online Appendix Table A.III.

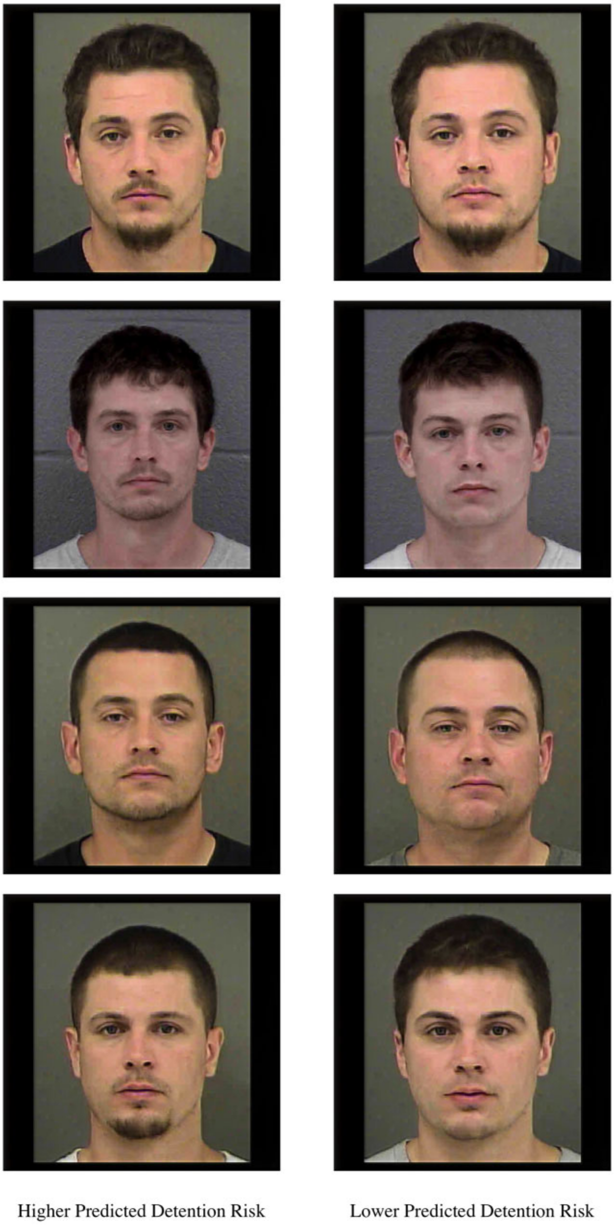
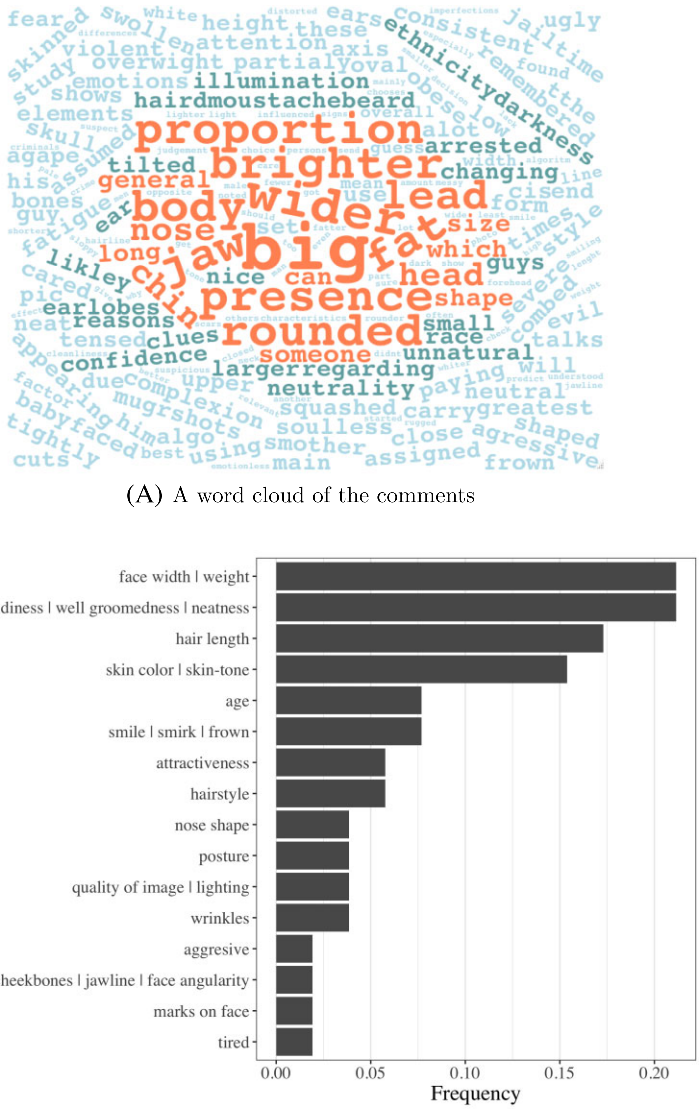

Higher Predicted Detention Risk
Lower Predicted Detention Risk

FIGURE IX

Examples of Morphing along the Orthogonal Gradients of the Face-Based Detention Predictor



(A) A word cloud of the comments

(B) Frequencies of comments by theme

FIGURE X

Subject Reports of What They See between Detention-Risk-Morphed Image Pairs, Orthogonalized to the First Novel Feature Discovered (Well-Groomed)

Panel A shows a word cloud of subject reports about what they see as the key difference between image pairs, where one is a randomly selected point in the

FIGURE X

(*Continued*) GAN latent space and the other is morphed in the direction of a higher predicted detention risk, where we are moving along the detention gradient orthogonal to well-groomed and skin tone (see the text). Panel B shows the frequency of semantic groupings of these open-ended subject reports (see the text for additional details).

"big," "wider," "presence," "rounded," "body," "jaw," and "head." When we ask independent research assistants to group the subject tokens into semantic groups, we can see as in the bottom of the figure that a sizable share of subject comments (around 22%) refer to a similar facial feature, $h_2(x)$: how "heavy-faced" or "full-faced" the defendant is.

This second facial feature (like the first) is again related to the algorithm's prediction of the judge. When we ask a separate sample of subjects (343 MTurk workers, see Online Appendix Table A.III) to independently label our validation images for heavy-facedness, we can see the R^2 from regressing the algorithm's predictions against heavy-faced yields an R^2 of 0.0384 (Table V, column (1)). With a coefficient of -0.0182 (0.0009), the results imply that a one standard deviation change in heavy-facedness (1.1946 points on our 9-point scale) is associated with a reduced predicted detention risk of 2.17 percentage points, or 9.3% of the base rate. Adding in other facial features implicated by past research substantially boosts the adjusted R^2 of the regression but barely changes the coefficient on heavy-facedness.

In principle, the procedure could be iterated further. After all, well-groomed, heavy-faced plus previously known facial features all together still only explain 27% of the variation in the algorithm's predictions of the judges' decisions. As long as there is residual variation, the hypothesis generation crank could be turned again and again. Because our goal is not to fully explain judges' decisions but to illustrate that the procedure works and is iterable, we leave this for future work (ideally done on data from other jurisdictions as well).

VI. EVALUATING THESE NEW HYPOTHESES

Here we consider whether the new hypotheses our procedure has generated meet our final criterion: empirical plausibility. We show that these facial features are new not just to the scientific literature but also apparently to criminal justice practitioners, before turning to whether these correlations might reflect some underlying causal relationship.

VI.A. Do These Hypotheses Predict What Judges Actually Do?

Empirical plausibility need not be implied by the fact that our new facial features are correlated with the algorithm's predictions of judges' decisions. The algorithm, after all, is not a perfect predictor. In principle, well-groomed and heavy-faced might be correlated with the part of the algorithm's prediction that is unrelated to judge behavior, or $m(x) - y$.

In Table VI, we show that our two new hypotheses are indeed empirically plausible. The adjusted R^2 from regressing judges' decisions against heavy-faced equals 0.0042 (column (1)), while for well-groomed the figure is 0.0021 (column (2)) and for both together the figure equals 0.0061 (column (3)). As a benchmark, the adjusted R^2 from all variables (other than the algorithm's overall mug shot–based prediction) in explaining judges' decisions equals 0.0218 (column (6)). So the explanatory power of our two novel hypotheses alone equals about 28% of what we get from all the variables together.

For a sense of the magnitude of these correlations, the coefficient on heavy-faced of −0.0234 (0.0036) in column (1) and on well-groomed of −0.0198 (0.0043) in column (2) imply that one standard deviation changes in each variable are associated with reduced detention rates equal to 2.8 and 2.0 percentage points, respectively, or 12.0% and 8.9% of the base rate. Interestingly, column (7) shows that heavy-faced remains statistically significant even when we control for the algorithm's prediction. The discovery procedure led us to a facial feature that, when measured independently, captures signal above and beyond what the algorithm found.[67]

VI.B. Do Practitioners Already Know This?

Our procedure has identified two hypotheses that are new to the existing research literature and to our study subjects. Yet the study subjects we have collected data from so far likely have relatively little experience with the criminal justice system. A reader might wonder: do experienced criminal justice practitioners already know that these "new" hypotheses affect judge decisions?

[67] See Online Appendix Figure A.XIX.

TABLE V

CORRELATION BETWEEN HEAVY-FACED AND THE ALGORITHM'S PREDICTION

	Dependent variable: Algorithmic judge detain prediction						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-faced	-0.0182^{***}	-0.0175^{***}	-0.0169^{***}	-0.0176^{***}	-0.0178^{***}	-0.0183^{***}	-0.0182^{***}
	(0.0009)	(0.0009)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Well-groomed		-0.0163^{***}	-0.0179^{***}	-0.0170^{***}	-0.0170^{***}	-0.0137^{***}	-0.0133^{***}
		(0.0011)	(0.0010)	(0.0010)	(0.0010)	(0.0012)	(0.0012)
Male			0.1193^{***}	0.1180^{***}	0.1152^{***}	0.1127^{***}	0.1129^{***}
			(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0024)
Age				0.0005^{***}	0.0005^{***}	0.0004^{***}	0.0004^{***}
				(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black				0.0057^{***}	-0.0179^{***}	-0.0181^{***}	-0.0183^{***}
				(0.0022)	(0.0035)	(0.0035)	(0.0035)
Asian				−0.0115	−0.0145	−0.0134	−0.0140
				(0.0111)	(0.0110)	(0.0110)	(0.0110)
Indigenous American				0.0078	0.0030	0.0046	0.0039
				(0.0232)	(0.0231)	(0.0230)	(0.0230)
Skin tone					-0.0488^{***}	-0.0469^{***}	-0.0472^{***}
					(0.0057)	(0.0057)	(0.0056)
Attractiveness						-0.0035^{**}	-0.0034^{**}
						(0.0016)	(0.0016)
Competence						-0.0062^{***}	-0.0061^{***}
						(0.0016)	(0.0016)

TABLE V

Continued

	Dependent variable: Algorithmic judge detain prediction						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dominance						0.0063^{***}	0.0058^{***}
						(0.0012)	(0.0012)
Trustworthiness						−0.0004	0.00003
						(0.0016)	(0.0016)
Human guess							0.0286^{***}
							(0.0060)
Constant	0.3485^{***}	0.4230^{***}	0.3340^{***}	0.3133^{***}	0.3568^{***}	0.3597^{***}	0.3423^{***}
	(0.0050)	(0.0070)	(0.0065)	(0.0073)	(0.0089)	(0.0098)	(0.0104)
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.0384	0.0603	0.2579	0.2613	0.2669	0.2711	0.2727

Notes. This table shows the results of estimating a linear probability specification regressing algorithmic predictions of judges' detain decision against different explanatory variables, using data from the validation set of cases from Mecklenburg County, NC. Each row of the table represents a different explanatory variable for the regression, while each column reports the results of a separate regression with different combinations of explanatory variables (as indicated by the filled-in coefficients and standard errors in the table). Algorithmic predictions of judges' decisions come from applying the algorithm built with face images in the training data set to validation set observations. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mug shot images (see the text). Human guess variable comes from showing subjects pairs of mug shot 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 and unknown gender. * $p < .1$; ** $p < .05$; *** $p < .01$.

TABLE VI

DO WELL-GROOMED AND HEAVY-FACED CORRELATE WITH JUDGE DECISIONS?

	Dependent variable: Judge detain decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-faced	−0.0234***		−0.0226***	−0.0223***		−0.0218***	−0.0111***
	(0.0036)		(0.0036)	(0.0036)		(0.0037)	(0.0037)
Well-groomed		−0.0198***	−0.0185***		−0.0124**	−0.0100*	−0.0022
		(0.0043)	(0.0043)		(0.0051)	(0.0051)	(0.0051)
Algo judge detain prediction							0.5842***
							(0.0449)
Male				0.0918***	0.0959***	0.0928***	0.0269**
				(0.0107)	(0.0108)	(0.0108)	(0.0118)
Age				−0.0011***	−0.0013***	−0.0012***	−0.0014***
				(0.0004)	(0.0004)	(0.0004)	(0.0004)
Black				−0.0645***	−0.0624***	−0.0643***	−0.0535***
				(0.0156)	(0.0156)	(0.0156)	(0.0154)
Asian				−0.0737	−0.0726	−0.0701	−0.0620
				(0.0488)	(0.0489)	(0.0488)	(0.0484)
Indigenous American				0.0490	0.0683	0.0524	0.0501
				(0.1019)	(0.1021)	(0.1019)	(0.1010)
Skin tone				−0.1062***	−0.1038***	−0.1076***	−0.0801***
				(0.0250)	(0.0251)	(0.0250)	(0.0249)
Attractiveness				−0.0084	0.0004	−0.0045	−0.0025
				(0.0067)	(0.0070)	(0.0070)	(0.0070)

TABLE VI

CONTINUED

	Dependent variable: Judge detain decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Competence				−0.0194***	−0.0175**	−0.0176**	−0.0141*
				(0.0072)	(0.0073)	(0.0073)	(0.0072)
Dominance				0.0109**	0.0076	0.0108**	0.0075
				(0.0052)	(0.0051)	(0.0052)	(0.0051)
Trustworthiness				−0.0085	−0.0104	−0.0075	−0.0075
				(0.0071)	(0.0071)	(0.0071)	(0.0070)
Human guess				0.1023***	0.1049***	0.0986***	0.0819***
				(0.0267)	(0.0268)	(0.0268)	(0.0266)
Constant	0.3569***	0.3280***	0.4418***	0.4436***	0.3642***	0.4665***	0.2666***
	(0.0196)	(0.0209)	(0.0276)	(0.0446)	(0.0429)	(0.0462)	(0.0483)
Naive-AUC	0.544	0.531	0.553	0.601	0.592	0.601	0.637
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R^2	0.0042	0.0021	0.0061	0.0215	0.0183	0.0218	0.0387

Notes. This table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables in the validation set described in Table I. The algorithmic predictions of the judges' detain decision come from our convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. Measures of defendant demographics and current arrest charge come from Mecklenburg County, NC, administrative data. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mug shot images (see the text). Human guess variable comes from showing subjects pairs of mug shot 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 and unknown gender. * $p < .1$; ** $p < .05$; *** $p < .01$.

The practitioners might have learned the influence of these facial features from day-to-day experience.

To answer this question, we carried out two smaller-scale data collections with a sample of $N = 15$ staff at a public defender's office and a legal aid society. We first asked an open-ended question: on what basis do judges decide to detain versus release defendants pretrial? Practitioners talked about judge misunderstandings of the law, people's prior criminal records, and judge underappreciation for the social contexts in which criminal records arise. Aside from the defendant's race, nothing about the appearance of defendants was mentioned.

We showed practitioners pairs of actual mug shots and asked them to guess which person is more likely to be detained by a judge (as we had done with MTurk and Prolific workers). This yields a sample of 360 detention forecasts. After seeing these mug shots practitioners were asked an open-ended question about what they think matters about the defendant's appearance for judge detention decisions. There were a few mentions of well-groomed and one mention of something related to heavy-faced, but these were far from the most frequently mentioned features, as seen in Online Appendix Figure A.XX.

The practitioner forecasts do indeed seem to be more accurate than those of "regular" study subjects. Table VII, column (5) shows that defendants whom the practitioners predict will be detained are 29.2 percentage points more likely to actually be detained, even after controlling for the other known determinants of detention from past research. This is nearly four times the effect of forecasts made by Prolific workers, as shown in the last column of Table VI. The practitioner guesses (unlike the regular study subjects) are even about as accurate as the algorithm; the R^2 from the practitioner guess (0.0165 in column (1)) is similar to the R^2 from the algorithm's predictions (0.0166 in column (6)).

Yet practitioners do not seem to already know what the algorithm has discovered. We can see this in several ways in Table VII. First, the sum of the adjusted R^2 values from the bivariate regressions of judge decisions against practitioner guesses and judge decisions against the algorithm mug shot–based prediction is not so different from the adjusted R^2 from including both variables in the same regression ($0.0165 + 0.0166 = 0.0331$ from columns (1) plus (6), versus 0.0338 in column (7)). We see something similar for the novel features of well-groomed and

TABLE VII

RESULTS FROM THE CRIMINAL JUSTICE PRACTITIONER SAMPLE

	Dependent variable: Judge detain decision							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Criminal justice practitioner guess	0.4172^{***}		0.3635^{**}		0.2924^{*}		0.4244^{***}	0.3395^{**}
	(0.1576)		(0.1592)		(0.1593)		(0.1562)	(0.1567)
Algo judge detain prediction						0.6201^{***}	0.6307^{***}	0.7555^{***}
						(0.2335)	(0.2315)	(0.2717)
Well-groomed		-0.0455^{*}	−0.0362	−0.0273	−0.0206			
		(0.0261)	(0.0263)	(0.0305)	(0.0306)			
Heavy-faced		-0.0394^{*}	-0.0363^{*}	-0.0411^{*}	-0.0387^{*}			
		(0.0217)	(0.0216)	(0.0217)	(0.0217)			
Male				−0.0696	−0.0680			-0.1579^{**}
				(0.0655)	(0.0653)			(0.0725)
Age				−0.0036	−0.0035			−0.0032
				(0.0029)	(0.0029)			(0.0028)
Black				-0.1683^{*}	-0.1706^{*}			−0.1454
				(0.0934)	(0.0931)			(0.0926)
Skin tone				-0.3901^{**}	-0.3895^{**}			-0.3192^{**}
				(0.1568)	(0.1562)			(0.1562)
Attractiveness				−0.0062	−0.0090			0.0049
				(0.0448)	(0.0447)			(0.0432)
Competence				0.0021	0.0039			0.0005
				(0.0441)	(0.0440)			(0.0434)

TABLE VII

CONTINUED

	Dependent variable: Judge detain decision							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dominance				0.0512*	0.0475			0.0334
				(0.0307)	(0.0307)			(0.0304)
Trustworthiness				−0.1113**	−0.1031**			−0.1145**
				(0.0446)	(0.0447)			(0.0443)
Constant	0.2855***	0.9205***	0.6778***	1.4446***	1.2442***	0.3377***	0.1226	0.7930***
	(0.0831)	(0.1662)	(0.1965)	(0.2728)	(0.2929)	(0.0646)	(0.1018)	(0.2679)
Naive-AUC	0.572	0.577	0.602	0.643	0.653	0.576	0.607	0.661
Observations	360	360	360	360	360	360	360	360
Adjusted R^2	0.0165	0.0131	0.0246	0.0384	0.0449	0.0166	0.0338	0.0582

Notes. This table shows the results of estimating judges' detain decisions using a linear probability specification of different explanatory variables on a subset of the validation set. The criminal justice practitioner's guess about the judge's decision comes from showing 15 different public defenders and legal aid society members actual mug shot images of defendants and asking them to report which defendant they believe the judge would be more likely to detain. The pairs are selected to be congruent in gender and race but discordant in detention outcome. The algorithmic predictions of judges' detain decisions come from applying the algorithm, which is 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, NC, administrative data. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mug shot images (see the text). Regression specifications also include indicators for unknown race and unknown gender. * $p < .1$; ** $p < .05$; *** $p < .01$.

heavy-faced specifically as well.[68] The practitioners and the algorithm seem to be tapping into largely unrelated signal.

VI.C. Exploring Causality

Are these novel features actually causally related to judge decisions? Fully answering that question is clearly beyond the scope of the present article. But we can present some additional evidence that is at least suggestive.

For starters we can rule out some obvious potential confounders. With the specific hypotheses in hand, identifying the most important concerns with confounding becomes much easier. In our application, well-groomed and heavy-faced could in principle be related to things like (say) the degree to which the defendant has a substance-abuse problem, is struggling with mental health, or their socioeconomic status. But as shown in a series of Online Appendix tables, we find that when we have study subjects independently label the mug shots in our validation data set for these features and then control for them, our novel hypotheses remain correlated with the algorithmic predictions of the judge and actual judge decisions.[69] We might wonder whether heavy-faced is simply a proxy for something that previous mock-trial-type studies suggest might matter for criminal justice decisions, "baby-faced" (Berry and Zebrowitz-McArthur 1988).[70] But when we have subjects rate mug shots for baby-facedness, our full-faced measure remains strongly predictive of the algorithm's

[68] The adjusted R^2 of including the practitioner forecasts plus well-groomed and heavy-facedness together (column (3), equal to 0.0246) is not that different from the sum of the R^2 values from including just the practitioner forecasts (0.0165 in column (1)) plus that from including just well-groomed and heavy-faced (equal to 0.0131 in Table VII, column (2)).

[69] In Online Appendix Table A.IX we show that controlling for one obvious indicator of a substance abuse issue—arrest for drugs—does not seem to substantially change the relationship between full-faced or well-groomed and the predicted detention decision. Online Appendix Tables A.X and A.XI show a qualitatively similar pattern of results for the defendant's mental health and socioeconomic status, which we measure by getting a separate sample of subjects to independently rate validation–data set mug shots. We see qualitatively similar results when the dependent variable is the actual rather than predicted judge decision; see Online Appendix Tables A.XIII, A.XIV, and A.XV.

[70] Characteristics of having a baby face included large eyes, narrow chin, small nose, and high, raised eyebrows. For a discussion of some of the larger literature on how that feature shapes the reactions of other people generally, see Zebrowitz et al. (2009).

predictions and actual judge decisions; see Online Appendix Tables A.XII and A.XVI.

In addition, we carried out a laboratory-style experiment with Prolific workers. We randomly morphed synthetic mug shot images in the direction of either higher or lower well-groomed (or full-faced), randomly assigned structured variables (current charge and prior record) to each image, explained to subjects the detention decision judges are asked to make, and then asked them which from each pair of subjects they would be more likely to detain if they were the judge. The framework from Mobius and Rosenblat (2006) helps clarify what this lab experiment gets us: appearance might affect how others treat us because others are reacting to something about our own appearance directly, because our appearance affects our own confidence, or because our appearance affects our effectiveness in oral communication. The experiment's results shut down these latter two mechanisms and isolate the effects of something about appearance per se, recognizing it remains possible well-groomed and heavy-faced are correlated with some other aspect of appearance.[71]

The study subjects recommend for detention those subjects with higher-risk structured variables (like current charge and prior record), which at the very least suggests they are taking the task seriously. Holding these other case characteristics constant, we find that the subjects are more likely to recommend for detention those defendants who are less well-groomed or less heavy-faced (see Online Appendix Table A.XVII). Qualitatively, these results support the idea that well-groomed and heavy-faced could have a causal effect. It is not clear that the magnitudes in these experiments necessarily have much meaning: the subjects are not actual judges, and the context and structure of choice is very different from real detention decisions. Still, it is worth noting that the magnitudes implied by our results are nontrivial. Changing well-groomed or heavy-faced has the same effect on subject decisions as a movement within the predicted rearrest risk distribution of 4 and 6 percentile points, respectively (see Online Appendix C for details). Of course only an actual field experiment could conclusively determine causality here, but carrying out that type of field experiment might seem more worthwhile to an investigator in light of the lab experiment's results.

[71] For additional details, see Online Appendix C.

Is this enough empirical support for these hypotheses to justify incurring the costs of causal testing? The empirical basis for these hypotheses would seem to be at least as strong as (or perhaps stronger than) the informal standard currently used to decide whether an idea is promising enough to test, which in our experience comes from some combination of observing the world, brainstorming, and perhaps some exploratory investigator-driven correlational analysis.

What might such causal testing look like? One possibility would follow in the spirit of Goldin and Rouse (2000) and compare detention decisions in settings where the defendant is more versus less visible to the judge to alter the salience of appearance. For example, many jurisdictions have continued to use some version of virtual hearings even after the pandemic.[72] In Chicago the court system has the defendant appear virtually but everyone else is in person, and the court system of its own volition has changed the size of the monitors used to display the defendant to court participants. One could imagine adding some planned variation to screen size or distance or angle to the judge. These video feeds could in principle be randomly selected for AI adjustment to the defendant's level of well-groomedness or heavy-facedness (this would probably fall into a legal gray area). In the case of well-groomed, one could imagine a field experiment that changed this aspect of the defendant's actual appearance prior to the court hearing. We are not claiming these are the right designs but intend only to illustrate that with new hypotheses in hand, economists are positioned to deploy the sort of creativity and rigorous testing that have become the hallmark of the field's efforts at causal inference.

VII. Conclusion

We have presented a new semi-automated procedure for hypothesis generation. We applied this new procedure to a concrete, socially important application: why judges jail some defendants and not others. Our procedure suggests two novel hypotheses: some defendants appear more well-groomed or more heavy-faced than others.

[72] See https://www.nolo.com/covid-19/virtual-criminal-court-appearances-in-the-time-of-the-covid-19.html.

Beyond the specific findings from our illustrative application, our empirical analysis also illustrates a playbook for other applications. Start with a high-dimensional predictor $m(x)$ of some behavior of interest. Build an unsupervised model of the data distribution, $p(x)$. Then combine the models for $m(x)$ and $p(x)$ in a morphing procedure to generate new instances that answer the counterfactual question: what would a given instance look like with higher or lower likelihood of the outcome? Show morphed pairs of instances to participants and get them to name what they see as the differences between morphed instances. Get others to independently rate instances for whatever the new hypothesis is; do these labels correlate with both $m(x)$ and the behavior of interest, y? If so, we have a new hypothesis worth causal testing. This playbook is broadly applicable whenever three conditions are met.

The first condition is that we have a behavior we can statistically predict. The application we examine here fits because the behavior is clearly defined and measured for many cases. A study of, say, human creativity would be more challenging because it is not clear that it can be measured (Said-Metwaly, Van den Noortgate, and Kyndt 2017). A study of why U.S. presidents use nuclear weapons during wartime would be challenging because there have been so few cases.

The second condition relates to what input data are available to predict behavior. Our procedure is likely to add only modest value in applications where we only have traditional structured variables, because those structured variables already make sense to people. Moreover the structured variables are usually already hypothesized to affect different behaviors, which is why economists ask about them on surveys. Our procedure will be more helpful with unstructured, high-dimensional data like images, language, and time series. The deeper point is that the collection of such high-dimensional data is often incidental to the scientific enterprise. We have images because the justice system photographs defendants during booking. Schools collect text from students as part of required assignments. Cellphones create location data as part of cell tower “pings.” These high-dimensional data implicitly contain an endless number of “features.”

Such high-dimensional data have already been found to predict outcomes in many economically relevant applications. Student essays predict graduation. Newspaper text predicts political slant of writers and editors. Federal Open Market Committee notes predict asset returns or volatility. X-ray images or

EKG results predict doctor diagnoses (or misdiagnoses). Satellite images predict the income or health of a place. Many more relationships like these remain to be explored. From such prediction models, one could readily imagine human inspection of morphs leading to novel features. For example, suppose high-frequency data on volume and stock prices are used to predict future excess returns, for example, to understand when the market over- or undervalues a stock. Morphs of these time series might lead us to discover the kinds of price paths that produce overreaction. After all, some investors have even named such patterns (e.g., “head and shoulders,” “double bottom”) and trade on them.

The final condition is to be able to morph the input data to create new cases that differ in the predicted outcome. This requires some unsupervised learning technique to model the data distribution. The good news is that a number of such techniques are now available that work well with different types of high-dimensional data. We happen to use GANs here because they work well with images. But our procedure can accomodate a variety of unsupervised models. For example for text we can use other methods like Bidirectional Encoder Representations from Transformers (Devlin et al. 2018), or for time series we could use variational auto-encoders (Kingma and Welling 2013).

An open question is the degree to which our experimental pipeline could be changed by new technologies, and in particular by recent innovations in generative modeling. For example, several recent models allow people to create new synthetic images from text descriptions, and so could perhaps (eventually) provide alternative approaches to the creation of counterfactual instances.[73] Similarly, recent generative language models appear to be able to process images (e.g., GPT-4), although they are only recently publicly available. Because there is inevitably some uncertainty in forecasting what those tools will be able to do in the future, they seem unlikely to be able to help with the first stage of our procedure’s pipeline—build a predictive model of some behavior of interest. To see why, notice that methods like GPT-4 are unlikely to have access to data on judge decisions linked to mug shots. But the stage of our pipeline that GPT-4 could potentially be helpful for is to substitute for humans in “naming” the contrasts between the morphed pairs of counterfactual instances.

[73] See https://stablediffusionweb.com/ and https://openai.com/product/dall-e-2.

Though speculative, such innovations potentially allow for more of the hypothesis generation procedure to be automated. We leave the exploration of these possibilities to future work.

Finally, it is worth emphasizing that hypothesis generation is not hypothesis testing. Each follows its own logic, and one procedure should not be expected to do both. Each requires different methods and approaches. What is needed to creatively produce new hypotheses is different from what is needed to carefully test a given hypothesis. Testing is about the curation of data, an effort to compare comparable subsets from the universe of all observations. But the carefully controlled experiment's focus on isolating the role of a single prespecified factor limits the ability to generate new hypotheses. Generation is instead about bringing as much data to bear as possible, since the algorithm can only consider signal within the data available to it. The more diverse the data sources, the more scope for discovery. An algorithm could have discovered judge decisions are influenced by football losses, as in Eren and Mocan (2018), but only if we thought to merge court records with massive archives of news stories as for example assembled by Leskovec, Backstrom, and Kleinberg (2009). For generating ideas, creativity in experimental design useful for testing is replaced with creativity in data assembly and merging.

More generally, we hope to raise interest in the curious asymmetry we began with. Idea generation need not remain such an idiosyncratic or nebulous process. Our framework hopefully illustrates that this process can also be modeled. Our results illustrate that such activity could bear actual empirical fruit. At a minimum, these results will hopefully spur more theoretical and empirical work on hypothesis generation rather than leave this as a largely "prescientific" activity.

University of Chicago and National Bureau of Economic Research, United States
University of Chicago and National Bureau of Economic Research, United States

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

Data Availability

The data underlying this article are available in the Harvard Dataverse, https://doi.org/10.7910/DVN/ILO46V (Ludwig and Mullainathan 2023b).

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 Children's Books," *Quarterly Journal of Economics*, 138 (2023), 2225–2285. https://doi.org/10.1093/qje/qjad028

Angelova, Victoria, Will S. Dobbie, and Crystal S. Yang, "Algorithmic Recommendations and Human Discretion," NBER Working Paper no. 31747, 2023. https://doi.org/10.3386/w31747

Arnold, David, Will S. Dobbie, and Peter Hull, "Measuring Racial Discrimination in Bail Decisions," NBER Working Paper no. 26999, 2020. https://doi.org/10.3386/w26999

Arnold, David, Will Dobbie, and Crystal S. Yang, "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133 (2018), 1885–1932. https://doi.org/10.1093/qje/qjy012

Athey, Susan, "Beyond Prediction: Using Big Data for Policy Problems," *Science*, 355 (2017), 483–485. https://doi.org/10.1126/science.aal4321

———, "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda,* Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds. (Chicago: University of Chicago Press, 2018), 507–547.

Athey, Susan, and Guido W. Imbens, "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11 (2019), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Athey, Susan, Guido W. Imbens, Jonas Metzger, and Evan Munro, "Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations," *Journal of Econometrics*, (2021), 105076. https://doi.org/10.1016/j.jeconom.2020.09.013

Athey, Susan, Dean Karlan, Emil Palikot, and Yuan Yuan, "Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces," NBER Working Paper no. 30633, 2022. https://doi.org/10.3386/w30633

Autor, David, "Polanyi's Paradox and the Shape of Employment Growth," NBER Working Paper no. 20485, 2014. https://doi.org/10.3386/w20485

Avitzour, Eliana, Adi Choen, Daphna Joel, and Victor Lavy, "On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes," NBER Working Paper no. 27818, 2020. https://doi.org/10.3386/w27818

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller, "How to Explain Individual Classification Decisions," *Journal of Machine Learning Research*, 11 (2010), 1803–1831.

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Begall, Sabine, Jaroslav Červený, Julia Neef, Oldřich Vojtěch, and Hynek Burda, "Magnetic Alignment in Grazing and Resting Cattle and Deer," *Proceedings of the National Academy of Sciences*, 105 (2008), 13451–13455. https://doi.org/10.1073/pnas.0803650105

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28 (2014), 29–50. https://doi.org/10.1257/jep.28.2.29

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*, 14 (1988), 23–33. https://doi.org/10.1177/0146167288141003

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*, 94 (2004), 991–1013. https://doi.org/10.1257/0002828042002561

Bjornstrom, Eileen E. S., 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*, 57 (2010), 269–293. https://doi.org/10.1525/sp.2010.57.2.269

Breiman, Leo, “Random Forests,” *Machine Learning*, 45 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, *Classification and Regression Trees* (London: Routledge, 1984). https://doi.org/10.1201/9781315139470

Brier, Glenn W., “Verification of Forecasts Expressed in Terms of Probability,” *Monthly Weather Review*, 78 (1950), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

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*, 91 (2019), 045002. https://doi.org/10.1103/RevModPhys.91.045002

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue, “Decision Making under the Gambler’s Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires,” *Quarterly Journal of Economics*, 131 (2016), 1181–1242. https://doi.org/10.1093/qje/qjw017

Chen, Daniel L., and Arnaud Philippe, “Clash of Norms: Judicial Leniency on Defendant Birthdays,” *Journal of Economic Behavior & Organization*, 211 (2023), 324–344. https://doi.org/10.1016/j.jebo.2023.05.002

Dahl, Gordon B., and Matthew M. Knepper, “Age Discrimination across the Business Cycle,” NBER Working Paper no. 27581, 2020. https://doi.org/10.3386/w27581

Davies, Alex, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli, “Advancing Mathematics by Guiding Human Intuition with AI,” *Nature*, 600 (2021), 70–74. https://doi.org/10.1038/s41586-021-04086-x

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. https://doi.org/10.48550/arXiv.1810.04805

Dobbie, Will, 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*, 108 (2018), 201–240. https://doi.org/10.1257/aer.20161503

Dobbie, Will, and Crystal S. Yang, “The US Pretrial System: Balancing Individual Rights and Public Interests,” *Journal of Economic Perspectives*, 35 (2021), 49–70. https://doi.org/10.1257/jep.35.4.49

Doshi-Velez, Finale, and Been Kim, “Towards a Rigorous Science of Interpretable Machine Learning,” arXiv preprint arXiv:1702.08608, 2017. https://doi.org/10.48550/arXiv.1702.08608

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*, 17 (2006), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Einav, Liran, and Jonathan Levin, “The Data Revolution and Economic Analysis,” *Innovation Policy and the Economy*, 14 (2014), 1–24. https://doi.org/10.1086/674019

Eren, Ozkan, and Naci Mocan, “Emotional Judges and Unlucky Juveniles,” *American Economic Journal: Applied Economics*, 10 (2018), 171–205. https://doi.org/10.1257/app.20160390

Frieze, Irene Hanson, Josephine E. Olson, and June Russell, “Attractiveness and Income for Men and Women in Management,” *Journal of Applied Social Psychology*, 21 (1991), 1039–1057. https://doi.org/10.1111/j.1559-1816.1991.tb00458.x

Fryer, Roland G., Jr, “An Empirical Analysis of Racial Differences in Police Use of Force: A Response,” *Journal of Political Economy*, 128 (2020), 4003–4008. https://doi.org/10.1086/710977

Fudenberg, Drew, and Annie Liang, “Predicting and Understanding Initial Play,” *American Economic Review*, 109 (2019), 4112–4141. https://doi.org/10.1257/aer.20180654

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, “Text as Data,” *Journal of Economic Literature*, 57 (2019), 535–574. https://doi.org/10.1257/jel.20181020

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 2022. https://doi.org/10.48550/arXiv.2105.15164

Goldin, Claudia, and Cecilia Rouse, “Orchestrating Impartiality: The Impact of ‘Blind’ Auditions on Female Musicians,” *American Economic Review*, 90 (2000), 715–741. https://doi.org/10.1257/aer.90.4.715

Goncalves, Felipe, and Steven Mello, “A Few Bad Apples? Racial Bias in Policing,” *American Economic Review*, 111 (2021), 1406–1441. https://doi.org/10.1257/aer.20181607

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*, 27 (2014), 2672–2680.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy, “Explaining and Harnessing Adversarial Examples,” arXiv preprint arXiv:1412.6572, 2014. https://doi.org/10.48550/arXiv.1412.6572

Grogger, Jeffrey, and Greg Ridgeway, “Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness,” *Journal of the American Statistical Association*, 101 (2006), 878–887. https://doi.org/10.1198/016214506000000168

Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2 (Berlin: Springer, 2009).

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*, 116 (2019), 13825–13832. https://doi.org/10.1073/pnas.1821458116

Heckman, James J., and Burton Singer, “Abducting Economics,” *American Economic Review*, 107 (2017), 298–302. https://doi.org/10.1257/aer.p20171118

Heyes, Anthony, and Soodeh Saberian, “Temperature and Decisions: Evidence from 207,000 Court Cases,” *American Economic Journal: Applied Economics*, 11 (2019), 238–265. https://doi.org/10.1257/app.20170223

Hoekstra, Mark, and CarlyWill Sloan, “Does Race Matter for Police Use of Force? Evidence from 911 Calls,” *American Economic Review*, 112 (2022), 827–860. https://doi.org/10.1257/aer.20201292

Hunter, Margaret, “The Persistent Problem of Colorism: Skin Tone, Status, and Inequality,” *Sociology Compass*, 1 (2007), 237–254. https://doi.org/10.1111/j.1751-9020.2007.00006.x

Jordan, Michael I., and Tom M. Mitchell, “Machine Learning: Trends, Perspectives, and Prospects,” *Science*, 349 (2015), 255–260. https://doi.org/10.1126/science.aaa8415

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and Anna Potapenko et al., “Highly Accurate Protein Structure

Prediction with AlphaFold," *Nature*, 596 (2021), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein, "Simple Rules for Complex Decisions," SSRN working paper, 2017. https://doi.org/10.2139/ssrn.2919024

Kahneman, Daniel, Olivier Sibony, and C. R Sunstein, *Noise* (London: HarperCollins, 2022).

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot, "An Adversarial Approach to Structural Estimation," University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2020-144, 2020. https://doi.org/10.2139/ssrn.3706365

Kingma, Diederik P., and Max Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013. https://doi.org/10.48550/arXiv.1312.6114

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133 (2018), 237–293. https://doi.org/10.1093/qje/qjx032

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*, 11 (2021), 10286. https://doi.org/10.1038/s41598-021-89743-x

Lahat, Dana, Tülay Adali, and Christian Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects," *Proceedings of the IEEE*, 103 (2015), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

Lang, Oran, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, and Michal Irani et al., "Explaining in Style: Training a GAN to Explain a Classifier in StyleSpace," paper presented at the IEEE/CVF International Conference on Computer Vision, 2021. https://doi.org/10.1109/ICCV48922.2021.00073

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg, "Meme-Tracking and the Dynamics of the News Cycle," paper presented at the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009. https://doi.org/10.1145/1557019.1557077

Little, Anthony C., Benedict C. Jones, and Lisa M. DeBruine, "Facial Attractiveness: Evolutionary Based Research," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366 (2011), 1638–1659. https://doi.org/10.1098/rstb.2010.0404

Liu, Shusen, Bhavya Kailkhura, Donald Loveland, and Yong Han, "Generative Counterfactual Introspection for Explainable Deep Learning," paper presented at the IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2019. https://doi.org/10.1109/GlobalSIP45357.2019.8969491

Ludwig, Jens, and Sendhil Mullainathan, "Machine Learning as a Tool for Hypothesis Generation," NBER Working Paper no. 31017, 2023a. https://doi.org/10.3386/w31017

———, "Replication Data for: 'Machine Learning as a Tool for Hypothesis Generation'," (2023b), Harvard Dataverse. https://doi.org/10.7910/DVN/ILO46V.

Marcinkevičs, Ričards, and Julia E. Vogt, "Interpretability and Explainability: A Machine Learning Zoo Mini-Tour," arXiv preprint arXiv:2012.01805, 2020. https://doi.org/10.48550/arXiv.2012.01805

Miller, Andrew, Ziad Obermeyer, John Cunningham, and Sendhil Mullainathan, "Discriminative Regularization for Latent Variable Models with Applications to Electrocardiography," paper presented at the International Conference on Machine Learning, 2019.

Mobius, Markus M., and Tanya S. Rosenblat, "Why Beauty Matters," *American Economic Review*, 96 (2006), 222–235. https://doi.org/10.1257/000282806776157515

Mobley, R. Keith, *An Introduction to Predictive Maintenance* (Amsterdam: Elsevier, 2002).

Mullainathan, Sendhil, and Ziad Obermeyer, “Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care,” *Quarterly Journal of Economics*, 137 (2022), 679–727. https://doi.org/10.1093/qje/qjab046

Mullainathan, Sendhil, and Jann Spiess, “Machine Learning: an Applied Econometric Approach,” *Journal of Economic Perspectives*, 31 (2017), 87–106. https://doi.org/10.1257/jep.31.2.87

Murphy, Allan H., “A New Vector Partition of the Probability Score,” *Journal of Applied Meteorology and Climatology*, 12 (1973), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

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. https://doi.org/10.48550/arXiv.1810.09136

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,* (Berlin: Springer, 2020), 273–283. https://doi.org/10.1007/978-3-030-59710-8_27

Neumark, David, Ian Burn, and Patrick Button, “Experimental Age Discrimination Evidence and the Heckman Critique,” *American Economic Review*, 106 (2016), 303–308. https://doi.org/10.1257/aer.p20161008

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*, 115 (2018), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

Oosterhof, Nikolaas N., and Alexander Todorov, “The Functional Basis of Face Evaluation,” *Proceedings of the National Academy of Sciences*, 105 (2008), 11087–11092. https://doi.org/10.1073/pnas.0805664105

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*, 372 (2021), 1209–1214. https://doi.org/10.1126/science.abe2629

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*, 27 (2021), 136–140. https://doi.org/10.1038/s41591-020-01192-7

Pion-Tonachini, Luca, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W. Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, and 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. https://doi.org/10.48550/arXiv.2111.13786

Popper, Karl, *The Logic of Scientific Discovery* (London: Routledge, 2nd ed. 2002). https://doi.org/10.4324/9780203994627

Pronin, Emily, “The Introspection Illusion,” *Advances in Experimental Social Psychology*, 41 (2009), 1–67. https://doi.org/10.1016/S0065-2601(08)00401-2

Ramachandram, Dhanesh, and Graham W. Taylor, “Deep Multimodal Learning: A Survey on Recent Advances and Trends,” *IEEE Signal Processing Magazine*, 34 (2017), 96–108. https://doi.org/10.1109/MSP.2017.2738401

Rambachan, Ashesh, “Identifying Prediction Mistakes in Observational Data,” Harvard University Working Paper, 2021. www.nber.org/system/files/chapters/c14777/c14777.pdf

Said-Metwaly, Sameh, Wim Van den Noortgate, and Eva Kyndt, "Approaches to Measuring Creativity: A Systematic Literature Review," *Creativity: Theories–Research-Applications*, 4 (2017), 238–275. https://doi.org/10.1515/ctra-2017-0013

Schickore, Jutta, "Scientific Discovery," in *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta, ed. (Stanford, CA: Stanford University, 2018).

Schlag, Pierre, "Law and Phrenology," *Harvard Law Review*, 110 (1997), 877–921. https://doi.org/10.2307/1342231

Sheetal, Abhishek, Zhiyu Feng, and Krishna Savani, "Using Machine Learning to Generate Novel Hypotheses: Increasing Optimism about COVID-19 Makes People Less Willing to Justify Unethical Behaviors," *Psychological Science*, 31 (2020), 1222–1235. https://doi.org/10.1177/0956797620959594

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," paper presented at the Workshop at International Conference on Learning Representations, 2014.

Sirovich, Lawrence, and Michael Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces," *Journal of the Optical Society of America A*, 4 (1987), 519–524. https://doi.org/10.1364/JOSAA.4.000519

Sunstein, Cass R., "Governing by Algorithm? No Noise and (Potentially) Less Bias," *Duke Law Journal*, 71 (2021), 1175–1205. https://doi.org/10.2139/ssrn.3925240

Swanson, Don R., "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge," *Perspectives in Biology and Medicine*, 30 (1986), 7–18. https://doi.org/10.1353/pbm.1986.0087

———, "Migraine and Magnesium: Eleven Neglected Connections," *Perspectives in Biology and Medicine*, 31 (1988), 526–557. https://doi.org/10.1353/pbm.1988.0009

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. https://doi.org/10.48550/arXiv.1312.6199

Todorov, Alexander, and DongWon Oh, "The Structure and Perceptual Basis of Social Judgments from Faces. in *Advances in Experimental Social Psychology*, B. Gawronski, ed. (Amsterdam: Elsevier, 2021), 189–245.

Todorov, Alexander, Christopher Y. Olivola, Ron Dotsch, and Peter Mende-Siedlecki, "Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance," *Annual Review of Psychology*, 66 (2015), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Varian, Hal R., "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28 (2014), 3–28. https://doi.org/10.1257/jep.28.2.3

Wilson, Timothy D., *Strangers to Ourselves* (Cambridge, MA: Harvard University Press, 2004).

Yuhas, Ben P., Moise H. Goldstein, and Terrence J. Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Communications Magazine*, 27 (1989), 65–71. https://doi.org/10.1109/35.41402

Zebrowitz, Leslie A., Victor X. Luevano, Philip M. Bronstad, and Itzhak Aharon, "Neural Activation to Babyfaced Men Matches Activation to Babies," *Social Neuroscience*, 4 (2009), 1–10. https://doi.org/10.1080/17470910701676236

Online Appendix

Machine Learning as a Tool for Hypothesis Generation

Jens Ludwig and Sendhil Mullainathan

A Appendix A: Data and Institutional Details

A.A. *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, Henderson, Miratrix and Valentine, 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.[74] 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, Doleac and Harvey (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 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.

A.B. 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.

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

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.

B Appendix B: 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.

B.A. 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.[75]

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

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

(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.[76] 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 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, Zhang, Ren and Sun (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

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

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[77] 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 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.

B.B. 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).[78] 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, Liu, Chen and Rudin (2018), Zhang, Wu and Zhu (2018), Ghorbani, Wexler, Zou and Kim (2019), and Chen, Li, Tao, Barnett, Rudin and Su (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

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

[78] 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.

present or not.[79] 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, Mittelstadt and Russell, 2018), or to assess the robustness of models, such as whether a breast cancer detector is looking in the right place (Bai, Wang, Liu, Liu, Song, Sebe and Kim, 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.[80] 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, the goal would be to generate an old face, and this is done when age is measured in the data during training.[81]

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.[82] 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 expla-

[79] 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.

[80] See, for example, Holte (1993), Rudin, Passonneau, Radeva, Dutta, Ierome and Isaac (2010), Freitas (2014), Letham, Rudin, McCormick and Madigan (2015), Angelino, Larus-Stone, Alabi, Seltzer and Rudin (2018), Jung et al. (2017), Chen and Rudin (2018), Ustun and Rudin (2019), Rudin (2019), and the references therein.

[81] 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.

[82] A recent ambitious paper has tried to tackle this problem. Peterson, Uddenberg, Griffiths, Todorov and Suchow (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."

nations"). 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, Creager, Goldenberg and Duvenaud, 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 Liu et al. (2019); Lang et al. (2021) and Ghandeharioun et al. (2022). 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., 2022). The ability of these alternative methods for generating counterfactual instances, or entirely new technologies that could be used for that task, we leave to future work. For our purposes the key point is that our own procedure appears to be capable of generating sufficiently high-quality morphed pairs of counterfactual instances to enable a sizable share of study subjects to articulate the same novel feature, which in turn is correlated both with the algorithm's predictions and actual judge decisions (as discussed in the text).

B.C. 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, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio (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 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, Laine and Aila (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, Ramsauer, Unterthiner, Nessler and Hochreiter, 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.[83] 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.[84] 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, Bachem, Lucic, Bousquet and Gelly, 2018), which are analogous to, but distinct from, common metrics of the same name used in 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, Laine, Aittala, Hellsten, Lehtinen and Aila, 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.

B.D. Morphing

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.

[83] 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.

[84] 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.

To produce a collection of morph pairs, we first fix a small positive constant α for the step size, and the constants $\check{m}$ and $\hat{m}$ 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 $\check{m} = 0.1$ and $\hat{m} = 0.35$, 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 $m(G(z^-)) \leq \check{m}$. 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 $m(G(z^+)) \geq \hat{m}$. 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.

B.D..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($g$, $m$, $h$, $\boldsymbol{z}$, $\alpha$, $y^{+}$)
     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
```

C Appendix C: 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 III), 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, Gelbach and Miller (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.[85] 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.[86]

[85] Recall the discussion in Section IV.B. argues against the possibility that these facial characteristics are proxies for risk.

[86] 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.

D Appendix D: Hold-out dataset results

As discussed in the main text, we downloaded data on $81,166$ arrests made between January 18, 2017, and January 17, 2020, involving $42,353$ unique defendants. We applied several data filters, such as dropping cases without mugshots (Appendix Table A.I.), which leaves us with $51,751$ observations. Because our goal is inference about *new*—that is, out-of-sample (OOS)–observations, we partitioned our data as follows:

- A *train set* of $N = 22,696$ cases, constructed by taking arrests through July 17, 2019, grouping arrests by arrestee,[87] randomly selecting 70% to the training-plus-validation dataset, then randomly selecting 70% of those arrestees for the training data specifically.
- A *validation set* of $N = 9,604$ cases used to report out-of-sample performance in this paper's main exhibits, consisting of the remaining 30% in the combined training-plus-validation data frame.
- A *hold-out set* of $N = 19,009$ cases that we did not touch until the paper was accepted for final publication, to avoid inadvertently overfitting the OOS data as we respond to seminar or referee suggestions, etc. This consists of the $N = 4,759$ valid cases for the last 6 months of our data period (July 17, 2019, to January 17, 2020) plus a random sample of 30% of those arrested before July 17, 2019.

The main exhibits in the paper report results from machine learning models built using the train set; results in the tables come from applying that machine learning model to observations in the validation dataset.

While this split-sample approach helps address concerns about over-fitting, there is nonetheless always a natural concern that we may still have inadvertently overfit the data over the course of writing the paper since we inevitably have built many machine learning models with the training dataset and evaluated their results with the validation dataset. To guard against that possibility, we replicated our analysis using the hold-out dataset, which we did not touch until this paper had been accepted for final publication.

The first step of this hold-out analysis was to collect new subject-assigned labels for the mugshot images in the hold-out dataset. We carried that out as follows:

- For subject guesses about detention likelihood (which of two mugshots will the judge be more likely to detain), our procedure for the hold-out dataset was exactly the same as for the validation dataset. We recruited $1,144$ participants from the Prolific platform and had them make guesses for a total of $56,688$ mugshot pairs ($35,110$ unique pairs). Participants were paid a \$1.80 base-rate to participate, plus an incentive of \$0.05 per correct guess. We collected an average of 5.6 guesses per image in the hold-out dataset.
- To measure the other explanatory variables not captured by administrative data (such as skin tone as well as facial features that previous psychological research suggests may shape the behavior of other people towards someone, like trustworthiness, dominance, competence and attractiveness) as well as the novel features that our discovery procedure identified (heavy-faced and well-groomed), we recruited $2,321$ subjects on Prolific. Subjects were asked to first label 50 images for the explanatory variables plus well-groomed and heavy-faced (with a payment of \$0.10 per image) then given another 50 images and asked to provide labels for just well-groomed and heavy-faced (with a

[87] We partition the data by arrestee, not arrest, to ensure people show up in only one of the partitions to avoid inadvertent information "leakage" across data partitions.

payment of \$0.05 per image). We collected an average of 3 labels per image for the explanatory variables (with a minimum of 2 labels per image) and an average of 5.7 labels per image for our two novel facial features (with a minimum of 5 labels per image).

The second step was to apply our convolutional neural network's (CNN) estimates to images in the hold-out dataset to calculate predicted detention likelihood for each image. One complication is that over the course of the project the code and model parameters of the original CNN we used for the main exhibits were lost. However we do have the predicted probabilities from that CNN for the validation dataset, and we built two new CNNs that each have identical explanatory power in the validation dataset as the original model (AUC of 0.6246). We use those new CNNs to calculate predicted detention probabilities for each image in the hold-out dataset for use as explanatory variables in the various regression exercises shown in the tables.

We then replicate the results in our main tables (Tables II through VI) three ways using our hold-out dataset:

- Using the entire hold-out dataset;
- Using just the part of the hold-out dataset that consists of the random sample of the 30% of people who had been arrested prior to July 17, 2019;
- using just the out-of-time partition of our hold-out dataset (the $N = 4,759$ valid cases for the last 6 months of our data period, from July 17, 2019, to January 17, 2020).

As shown in Appendix Tables A.XVIII to A.XXXII, our core findings about the influence of facial features on judge decisions and which specific facial features matter most are qualitatively similar to those shown in the main tables.

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.
Missing charge flags	(18)	Rows for which we are missing a categorization of charge descriptions.
Prisoner level separation	(2712)	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	
Sample Partitioning		
Train Set	22696	This is the set on which we trained our judge prediction algorithm.
Validation Set	10046	We use this set to report out-of-sample performance in this paper's main exhibits.[88]
Lock-box Hold-Out Set	19009	The data we have set aside for measuring the model's final performance, composed of a combined out-of-sample by individual subset and an out-of-sample by time subset.

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

[88] An additional 442 observations were removed from the validation set after our CNN failed to generate predictions for these mugshots, so results on the validation set are reported on a sample size of 9604 throughout this paper.

Table A.II: Test of balance between training dataset, validation dataset and out-of-sample by individual testing dataset

	Train Set	**Validation Set**	**Lock-Box Hold-Out Data (OOS by individual)**	**Comparison p-value**
Sample Size	22696	9604	14250	
Outcome				
Judge detains defendant	0.232	0.233	0.235	0.883
Defendant re-arrested before trial	0.251	0.251	0.255	0.876
Defendant Characteristics				
Age	31.849	31.631	32.171	0.110
Male	0.789	0.782	0.778	0.245
White	0.278	0.274	0.285	0.478
Black	0.694	0.695	0.687	0.686
Other Race	0.028	0.031	0.027	
Arrest Charge				
Violent	0.343	0.343	0.339	0.727
Property	0.324	0.317	0.319	0.614
Drug	0.204	0.207	0.198	0.403
Gun	0.079	0.084	0.078	0.352
Other charge	0.262	0.264	0.272	0.155
Arrest Charge Severity				
Felony	0.421	0.428	0.410	0.073*
Non-Felony	0.579	0.572	0.590	
Defendant Prior Record				
Any Prior Conviction	0.461	0.458	0.452	0.575
Prior Felony Conviction	0.333	0.328	0.323	0.569
Prior Non-Felony Conviction	0.316	0.318	0.313	0.898

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 lock-box hold-out hold-out data set (OOS by individual) consists of data from a subset of cases through July 16, 2019, selected by randomly selecting arrestees, excluding the last 6 months of our study period (July 17, 2019, through January 17, 2019), which is kept for the lock-box hold-out data (OOS by time). 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 (which we use to report results in our paper's main exhibits). 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 comparison p-value comes from calculating an F-test statistic for the null hypothesis of equivalence of means for a given variable (described by each row label) between the training data set, validation data set and the lock-box hold-out data set (OOS by individual only), with standard errors clustered by arrestee.

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 $\geq$ 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 subject's 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

Table A.XVIII: Is the algorithm rediscovering known facial features? (complete lock-box hold-out)

	Dependent variable:				
	Algo Judge Detain Prediction				
	(1)	(2)	(3)	(4)	(5)
Male	0.0734***	0.0723***	0.0713***	0.0710***	0.0708***
	(0.0013)	(0.0013)	(0.0013)	(0.0014)	(0.0014)
Age		−0.0002***	−0.0002***	−0.0002***	−0.0002***
		(0.00005)	(0.00005)	(0.0001)	(0.0001)
Black		0.0207***	0.0138***	0.0143***	0.0140***
		(0.0012)	(0.0015)	(0.0015)	(0.0015)
Asian		−0.0023	−0.0053	−0.0029	−0.0029
		(0.0064)	(0.0064)	(0.0063)	(0.0063)
Indigenous American		−0.0024	−0.0051	−0.0055	−0.0046
		(0.0146)	(0.0146)	(0.0145)	(0.0145)
Skin-Tone			−0.0240***	−0.0239***	−0.0242***
			(0.0031)	(0.0031)	(0.0031)
Attractiveness				0.0004	0.0005
				(0.0005)	(0.0005)
Competence				−0.0037***	−0.0036***
				(0.0007)	(0.0007)
Dominance				−0.0004	−0.0006
				(0.0005)	(0.0005)
Trustworthiness				−0.0047***	−0.0044***
				(0.0006)	(0.0006)
Human Guess					0.0257***
					(0.0033)
Constant	0.1508***	0.1436***	0.1630***	0.1975***	0.1838***
	(0.0012)	(0.0021)	(0.0033)	(0.0042)	(0.0046)
Observations	19,009	19,009	19,009	19,009	19,009
Adjusted R^2	0.1360	0.1506	0.1533	0.1656	0.1681

Notes: This table replicates the analysis from Table II but applies it to the complete lock-box hold-out data set. The table presents the results of regressing an algorithmic prediction of judge detention decisions against each of the different explanatory variables as listed in the rows, where each column represents a different regression specification. The algorithm was trained using mugshots from the training data set, and evaluated on pooled hold-out data including both in-time (randomly selected arrested prior to July 17, 2019) and out-of-time partitions (valid arrests from the last 6 months of the data period). Data on skin tone, attractiveness, competence, dominance, and trustworthiness comes from asking subjects to assign feature ratings to mugshot images from the Mecklenburg County, NC Sheriff's Office public website. The human guess about the judges' decision comes from showing workers on the Prolific platform pairs of mugshot images and asking them to report which defendant they believe the judge would be more likely to detain. Regressions follow a linear probability model and also include indicators for unknown race and unknown gender.

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

Table A.XIX: Does algorithm predict judge behavior after controlling for known factors? (complete lock-box hold-out)

	Dependent variable:						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algo Judge Detain Prediction	0.9226^{***}					0.8761^{***}	0.8756^{***}
	(0.0349)					(0.0381)	(0.0382)
Male		0.0910^{***}	0.0908^{***}		0.0877^{***}	0.0255^{***}	0.0255^{***}
		(0.0071)	(0.0071)		(0.0072)	(0.0076)	(0.0076)
Age		-0.0003	-0.0003		-0.0006^{**}	-0.0005^{*}	-0.0005^{*}
		(0.0003)	(0.0003)		(0.0003)	(0.0003)	(0.0003)
Black		-0.0177^{***}	-0.0192^{**}		-0.0183^{**}	-0.0308^{***}	-0.0308^{***}
		(0.0066)	(0.0081)		(0.0081)	(0.0080)	(0.0080)
Asian		-0.1173^{***}	-0.1179^{***}		-0.1128^{***}	-0.1103^{***}	-0.1103^{***}
		(0.0337)	(0.0337)		(0.0337)	(0.0332)	(0.0332)
Indigenous American		-0.0038	-0.0044		-0.0044	0.0004	0.0006
		(0.0774)	(0.0774)		(0.0773)	(0.0762)	(0.0762)
Skin-Tone			-0.0053		-0.0031	0.0179	0.0178
			(0.0163)		(0.0163)	(0.0161)	(0.0161)
Attractiveness				-0.0079^{***}	-0.0064^{**}	-0.0068^{**}	-0.0068^{**}
				(0.0026)	(0.0028)	(0.0028)	(0.0028)
Competence				-0.0065^{*}	-0.0087^{**}	-0.0054	-0.0054
				(0.0035)	(0.0036)	(0.0035)	(0.0035)
Dominance				0.0043^{*}	0.0003	0.0007	0.0006
				(0.0024)	(0.0024)	(0.0024)	(0.0024)
Trustworthiness				-0.0108^{***}	-0.0080^{**}	-0.0039	-0.0039
				(0.0033)	(0.0033)	(0.0033)	(0.0033)
Human Guess							0.0044
							(0.0176)
Constant	0.0230^{***}	0.1677^{***}	0.1720^{***}	0.2911^{***}	0.2722^{***}	0.0992^{***}	0.0969^{***}
	(0.0078)	(0.0113)	(0.0174)	(0.0142)	(0.0224)	(0.0233)	(0.0250)
Naive-AUC	0.63	0.559	0.558	0.546	0.576	0.637	0.637
Observations	19,009	19,009	19,009	19,009	19,009	19,009	19,009
Adjusted R^2	0.0354	0.0092	0.0092	0.0045	0.0129	0.0396	0.0395

Notes: This table replicates the analysis from Table III but applies it to the complete lock-box hold-out data set. The table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables, including both the in-time (randomly selected arrested prior to July 17, 2019) and out-of-time partitions (valid arrests from the last 6 months of the data period). The algorithmic predictions of the judges' detain decision come from our convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. Measures of defendant demographics and current arrest charge come from government administrative data obtained from a combination of Mecklenburg County, NC and state agencies. Measures of skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XX: Correlation between well-groomed (first novel feature) and algorithm's prediction (complete lock-box hold-out)

	Dependent variable:					
	Algo Judge Detain Prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Well-Groomed	-0.0073^{***}	-0.0076^{***}	-0.0085^{***}	-0.0085^{***}	-0.0067^{***}	-0.0065^{***}
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0006)	(0.0006)
Male		0.0737^{***}	0.0728^{***}	0.0718^{***}	0.0718^{***}	0.0716^{***}
		(0.0013)	(0.0013)	(0.0013)	(0.0014)	(0.0014)
Age			-0.0004^{***}	-0.0004^{***}	-0.0003^{***}	-0.0003^{***}
			(0.00005)	(0.00005)	(0.0001)	(0.0001)
Black			0.0203^{***}	0.0134^{***}	0.0138^{***}	0.0135^{***}
			(0.0012)	(0.0015)	(0.0015)	(0.0015)
Asian			−0.0001	−0.0030	−0.0018	−0.0019
			(0.0063)	(0.0063)	(0.0063)	(0.0063)
Indigenous American			−0.0055	−0.0082	−0.0081	−0.0072
			(0.0145)	(0.0145)	(0.0144)	(0.0144)
Skin-Tone				-0.0241^{***}	-0.0244^{***}	-0.0247^{***}
				(0.0031)	(0.0031)	(0.0030)
Attractiveness					0.0022^{***}	0.0022^{***}
					(0.0005)	(0.0005)
Competence					-0.0024^{***}	-0.0023^{***}
					(0.0007)	(0.0007)
Dominance					−0.0001	−0.0003
					(0.0005)	(0.0005)
Trustworthiness					-0.0036^{***}	-0.0035^{***}
					(0.0006)	(0.0006)
Human Guess						0.0232^{***}
						(0.0033)
Constant	0.2392^{***}	0.1839^{***}	0.1870^{***}	0.2065^{***}	0.2123^{***}	0.1993^{***}
	(0.0022)	(0.0023)	(0.0032)	(0.0040)	(0.0044)	(0.0047)
Observations	19,009	19,009	19,009	19,009	19,009	19,009
Adjusted R^2	0.0117	0.1487	0.1654	0.1681	0.1716	0.1737

Notes: This table replicates the analysis from Table IV but applies it to the complete lock-box hold-out data set, including both in-time (randomly selected arrested prior to July 17, 2019) and out-of-time partitions (valid arrests from the last 6 months of the data period). The table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to hold-out set observations. Data on well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXI: Correlation between heavy-faced (second novel feature) and algorithm's prediction (complete lock-box hold-out)

	Dependent variable:						
	Algo Judge Detain Prediction						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	−0.0145***	−0.0138***	−0.0131***	−0.0135***	−0.0137***	−0.0139***	−0.0138***
	(0.0004)	(0.0005)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Well-Groomed		−0.0051***	−0.0056***	−0.0063***	−0.0062***	−0.0045***	−0.0043***
		(0.0005)	(0.0004)	(0.0005)	(0.0005)	(0.0006)	(0.0006)
Male			0.0725***	0.0713***	0.0701***	0.0687***	0.0685***
			(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)
Age				−0.0003***	−0.0003***	−0.0003***	−0.0003***
				(0.00005)	(0.00005)	(0.00005)	(0.00005)
Black				0.0230***	0.0147***	0.0145***	0.0143***
				(0.0012)	(0.0015)	(0.0015)	(0.0015)
Asian				0.0056	0.0021	0.0031	0.0029
				(0.0061)	(0.0061)	(0.0061)	(0.0061)
Indigenous American				0.0016	−0.0015	−0.0009	−0.0003
				(0.0141)	(0.0141)	(0.0141)	(0.0141)
Skin-Tone					−0.0288***	−0.0279***	−0.0281***
					(0.0030)	(0.0030)	(0.0030)
Attractiveness						−0.0008	−0.0008
						(0.0005)	(0.0005)
Competence						−0.0024***	−0.0023***
						(0.0007)	(0.0007)
Dominance						0.0021***	0.0019***
						(0.0004)	(0.0004)
Trustworthiness						−0.0009	−0.0007
						(0.0006)	(0.0006)
Human Guess							0.0191***
							(0.0032)
Constant	0.2807***	0.2994***	0.2420***	0.2433***	0.2675***	0.2674***	0.2563***
	(0.0023)	(0.0029)	(0.0029)	(0.0035)	(0.0043)	(0.0046)	(0.0049)
Observations	19,009	19,009	19,009	19,009	19,009	19,009	19,009
Adjusted R^2	0.0522	0.0577	0.1903	0.2092	0.2130	0.2153	0.2167

Notes: This table replicates the analysis from Table V but applies it to the complete lock-box hold-out data set, including both in-time (randomly selected arrested prior to July 17, 2019) and out-of-time partitions (valid arrests from the last 6 months of the data period). The table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to hold-out set observations. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXII: Do well-groomed and heavy-faced (first and second novel features) correlate with judge decisions? (complete lock-box hold-out)

	Dependent variable:						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	−0.0257***		−0.0229***	−0.0242***		−0.0226***	−0.0113***
	(0.0023)		(0.0023)	(0.0023)		(0.0023)	(0.0024)
Well-Groomed		−0.0238***	−0.0202***		−0.0208***	−0.0172***	−0.0137***
		(0.0024)	(0.0024)		(0.0031)	(0.0031)	(0.0030)
Algo Judge Detain Prediction							0.8168***
							(0.0393)
Male				0.0825***	0.0901***	0.0851***	0.0291***
				(0.0072)	(0.0072)	(0.0072)	(0.0077)
Age				−0.0008***	−0.0009***	−0.0010***	−0.0007***
				(0.0003)	(0.0003)	(0.0003)	(0.0003)
Black				−0.0175**	−0.0201**	−0.0188**	−0.0305***
				(0.0081)	(0.0081)	(0.0081)	(0.0080)
Asian				−0.1038***	−0.1094***	−0.1016***	−0.1040***
				(0.0336)	(0.0337)	(0.0336)	(0.0332)
Indigenous American				0.0072	−0.0118	−0.0005	−0.0002
				(0.0770)	(0.0772)	(0.0770)	(0.0761)
Skin-Tone				−0.0096	−0.0049	−0.0104	0.0125
				(0.0163)	(0.0163)	(0.0163)	(0.0161)
Attractiveness				−0.0106***	−0.0010	−0.0058**	−0.0052*
				(0.0029)	(0.0029)	(0.0030)	(0.0029)
Competence				−0.0078**	−0.0045	−0.0045	−0.0026
				(0.0035)	(0.0036)	(0.0036)	(0.0035)
Dominance				0.0042*	0.0011	0.0047*	0.0031
				(0.0025)	(0.0024)	(0.0025)	(0.0024)
Trustworthiness				−0.0024	−0.0047	−0.0002	0.0004
				(0.0034)	(0.0034)	(0.0034)	(0.0034)
Human Guess				0.0183	0.0188	0.0122	−0.0034
				(0.0177)	(0.0178)	(0.0178)	(0.0176)
Constant	0.3444***	0.3180***	0.4180***	0.3671***	0.3077***	0.4011***	0.1918***
	(0.0118)	(0.0109)	(0.0147)	(0.0264)	(0.0253)	(0.0271)	(0.0286)
Naive-AUC	0.558	0.549	0.572	0.595	0.584	0.6	0.643
Observations	19,009	19,009	19,009	19,009	19,009	19,009	19,009
Adjusted R^2	0.0068	0.0051	0.0103	0.0186	0.0154	0.0202	0.0419

Notes: This table replicates the analysis from Table VI but applies it to the complete lock-box hold-out data set, including both in-time (randomly selected arrested prior to July 17, 2019) and out-of-time partitions (valid arrests from the last 6 months of the data period). The table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables. The algorithmic predictions of the judges' detain decision come from a convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. 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. 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 and unknown gender.

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

Table A.XXIII: Is the algorithm rediscovering known facial features? (lock-box hold-out data, OOS by individual)

	Dependent variable:				
	Algo Judge Detain Prediction				
	(1)	(2)	(3)	(4)	(5)
Male	0.0761***	0.0747***	0.0735***	0.0728***	0.0726***
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Age		−0.0001**	−0.0001**	−0.0002**	−0.0001**
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		0.0211***	0.0120***	0.0127***	0.0125***
		(0.0015)	(0.0018)	(0.0018)	(0.0018)
Asian		−0.0030	−0.0069	−0.0034	−0.0036
		(0.0079)	(0.0079)	(0.0078)	(0.0078)
Indigenous American		0.0027	−0.0020	−0.0008	−0.0005
		(0.0185)	(0.0185)	(0.0183)	(0.0183)
Skin-Tone			−0.0332***	−0.0328***	−0.0331***
			(0.0037)	(0.0037)	(0.0037)
Attractiveness				0.0007	0.0008
				(0.0006)	(0.0006)
Competence				−0.0039***	−0.0037***
				(0.0008)	(0.0008)
Dominance				−0.0001	−0.0003
				(0.0005)	(0.0005)
Trustworthiness				−0.0054***	−0.0051***
				(0.0008)	(0.0008)
Human Guess					0.0274***
					(0.0040)
Constant	0.1589***	0.1495***	0.1762***	0.2121***	0.1972***
	(0.0014)	(0.0026)	(0.0039)	(0.0050)	(0.0055)
Observations	14,250	14,250	14,250	14,250	14,250
Adjusted R^2	0.1320	0.1453	0.1501	0.1634	0.1660

Notes: This table replicates the analysis from Table II but applies it to the in-time hold-out data. The table presents the results of regressing an algorithmic prediction of judge detention decisions against each of the different explanatory variables as listed in the rows, where each column represents a different regression specification. The algorithm was trained using mugshots from the training data set, and evaluated on in-time (randomly selected arrested prior to July 17, 2019) partition of the hold-out set. Data on skin tone, attractiveness, competence, dominance, and trustworthiness comes from asking subjects to assign feature ratings to mugshot images from the Mecklenburg County, NC Sheriff's Office public website. The human guess about the judges' decision comes from showing workers on the Prolific platform pairs of mugshot images and asking them to report which defendant they believe the judge would be more likely to detain. Regressions follow a linear probability model and also include indicators for unknown race and unknown gender.

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

Table A.XXIV: Does algorithm predict judge behavior after controlling for known factors? (lock-box hold-out data, OOS by individual)

	Dependent variable:						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algo Judge Detain Prediction	0.8635***					0.8170***	0.8176***
	(0.0402)					(0.0438)	(0.0439)
Male		0.0923***	0.0917***		0.0879***	0.0284***	0.0284***
		(0.0085)	(0.0086)		(0.0087)	(0.0092)	(0.0092)
Age		-0.0005	-0.0005		-0.0008^{**}	-0.0007^{**}	-0.0007^{**}
		(0.0003)	(0.0003)		(0.0003)	(0.0003)	(0.0003)
Black		-0.0259^{***}	-0.0305^{***}		-0.0294^{***}	-0.0397^{***}	-0.0397^{***}
		(0.0079)	(0.0096)		(0.0096)	(0.0095)	(0.0095)
Asian		-0.1457^{***}	-0.1476^{***}		-0.1396^{***}	-0.1368^{***}	-0.1368^{***}
		(0.0413)	(0.0414)		(0.0414)	(0.0409)	(0.0409)
Indigenous American		0.0134	0.0111		0.0167	0.0173	0.0173
		(0.0971)	(0.0971)		(0.0970)	(0.0958)	(0.0958)
Skin-Tone			-0.0167		-0.0135	0.0132	0.0133
			(0.0194)		(0.0194)	(0.0193)	(0.0193)
Attractiveness				-0.0066^{**}	-0.0060^{*}	-0.0066^{**}	-0.0066^{**}
				(0.0031)	(0.0034)	(0.0033)	(0.0033)
Competence				-0.0065	-0.0084^{**}	-0.0053	-0.0053
				(0.0042)	(0.0042)	(0.0042)	(0.0042)
Dominance				0.0041	0.0008	0.0009	0.0010
				(0.0029)	(0.0029)	(0.0029)	(0.0029)
Trustworthiness				-0.0125^{***}	-0.0094^{**}	-0.0050	-0.0051
				(0.0040)	(0.0040)	(0.0040)	(0.0040)
Human Guess							-0.0050
							(0.0212)
Constant	0.0470***	0.1964***	0.2098***	0.3148***	0.3102***	0.1369***	0.1395***
	(0.0094)	(0.0136)	(0.0207)	(0.0169)	(0.0265)	(0.0278)	(0.0299)
Naive-AUC	0.618	0.561	0.558	0.544	0.573	0.628	0.628
Observations	14,250	14,250	14,250	14,250	14,250	14,250	14,250
Adjusted R^2	0.0313	0.0093	0.0093	0.0043	0.0129	0.0364	0.0363

Notes: This table replicates the analysis from Table III but applies it to the in-time hold-out data. The table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables in the in-time (randomly selected arrested prior to July 17, 2019) partition of the hold-out set. The algorithmic predictions of the judges' detain decision come from our convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. Measures of defendant demographics and current arrest charge come from government administrative data obtained from a combination of Mecklenburg County, NC and state agencies. Measures of skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXV: Correlation between well-groomed (first novel feature) and algorithm's prediction (lock-box hold-out data, OOS by individual)

	Dependent variable:					
	Algo Judge Detain Prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Well-Groomed	-0.0072^{***}	-0.0076^{***}	-0.0083^{***}	-0.0083^{***}	-0.0061^{***}	-0.0058^{***}
	(0.0006)	(0.0005)	(0.0006)	(0.0006)	(0.0007)	(0.0007)
Male		0.0764^{***}	0.0753^{***}	0.0740^{***}	0.0738^{***}	0.0735^{***}
		(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Age			-0.0003^{***}	-0.0003^{***}	-0.0002^{***}	-0.0002^{***}
			(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black			0.0209^{***}	0.0118^{***}	0.0123^{***}	0.0121^{***}
			(0.0015)	(0.0018)	(0.0018)	(0.0018)
Asian			-0.0006	-0.0044	-0.0026	-0.0028
			(0.0078)	(0.0078)	(0.0078)	(0.0078)
Indigenous American			0.0015	-0.0032	-0.0028	-0.0025
			(0.0184)	(0.0183)	(0.0183)	(0.0182)
Skin-Tone				-0.0330^{***}	-0.0332^{***}	-0.0334^{***}
				(0.0037)	(0.0037)	(0.0037)
Attractiveness					0.0023^{***}	0.0023^{***}
					(0.0007)	(0.0007)
Competence					-0.0028^{***}	-0.0026^{***}
					(0.0008)	(0.0008)
Dominance					0.0001	-0.0001
					(0.0005)	(0.0005)
Trustworthiness					-0.0044^{***}	-0.0042^{***}
					(0.0008)	(0.0008)
Human Guess						0.0252^{***}
						(0.0040)
Constant	0.2493^{***}	0.1917^{***}	0.1919^{***}	0.2183^{***}	0.2250^{***}	0.2107^{***}
	(0.0026)	(0.0027)	(0.0038)	(0.0048)	(0.0052)	(0.0057)
Observations	14,250	14,250	14,250	14,250	14,250	14,250
Adjusted R^2	0.0104	0.1437	0.1586	0.1633	0.1679	0.1701

Notes: This table replicates the analysis from Table IV but applies it to the in-time hold-out data, consisting of randomly selected cases arrested prior to July 17, 2019. The table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to hold-out set observations. Data on well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXVI: Correlation between heavy-faced (second novel feature) and algorithm's prediction (lock-box hold-out data, OOS by individual)

	Dependent variable:						
	Algo Judge Detain Prediction						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	-0.0149^{***}	-0.0142^{***}	-0.0136^{***}	-0.0142^{***}	-0.0145^{***}	-0.0147^{***}	-0.0146^{***}
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Well-Groomed		-0.0049^{***}	-0.0054^{***}	-0.0060^{***}	-0.0059^{***}	-0.0037^{***}	-0.0035^{***}
		(0.0006)	(0.0005)	(0.0005)	(0.0005)	(0.0007)	(0.0007)
Male			0.0754^{***}	0.0739^{***}	0.0724^{***}	0.0707^{***}	0.0705^{***}
			(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Age				-0.0003^{***}	-0.0003^{***}	-0.0003^{***}	-0.0003^{***}
				(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black				0.0245^{***}	0.0142^{***}	0.0139^{***}	0.0138^{***}
				(0.0015)	(0.0018)	(0.0018)	(0.0018)
Asian				0.0060	0.0017	0.0033	0.0032
				(0.0076)	(0.0076)	(0.0076)	(0.0076)
Indigenous American				0.0101	0.0049	0.0056	0.0058
				(0.0179)	(0.0178)	(0.0178)	(0.0178)
Skin-Tone					-0.0379^{***}	-0.0369^{***}	-0.0371^{***}
					(0.0036)	(0.0036)	(0.0036)
Attractiveness						-0.0007	-0.0007
						(0.0007)	(0.0007)
Competence						-0.0028^{***}	-0.0028^{***}
						(0.0008)	(0.0008)
Dominance						0.0025^{***}	0.0023^{***}
						(0.0005)	(0.0005)
Trustworthiness						-0.0015^{**}	-0.0014^{*}
						(0.0008)	(0.0008)
Human Guess							0.0212^{***}
							(0.0039)
Constant	0.2929^{***}	0.3108^{***}	0.2515^{***}	0.2497^{***}	0.2810^{***}	0.2813^{***}	0.2689^{***}
	(0.0028)	(0.0035)	(0.0035)	(0.0042)	(0.0052)	(0.0054)	(0.0059)
Observations	14,250	14,250	14,250	14,250	14,250	14,250	14,250
Adjusted R^2	0.0505	0.0552	0.1849	0.2030	0.2092	0.2125	0.2140

Notes: This table replicates the analysis from Table V but applies it to the in-time hold-out data, consisting of randomly selected cases arrested prior to July 17, 2019. The table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to hold-out set observations. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXVII: Do well-groomed and heavy-faced (first and second novel features) correlate with judge decisions? (lock-box hold-out data, OOS by individual)

	Dependent variable:						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	−0.0240***		−0.0212***	−0.0223***		−0.0207***	−0.0095***
	(0.0027)		(0.0027)	(0.0028)		(0.0028)	(0.0029)
Well-Groomed		−0.0233***	−0.0200***		−0.0206***	−0.0173***	−0.0146***
		(0.0029)	(0.0029)		(0.0036)	(0.0037)	(0.0036)
Algo Judge Detain Prediction							0.7693***
							(0.0452)
Male				0.0837***	0.0909***	0.0866***	0.0324***
				(0.0087)	(0.0087)	(0.0087)	(0.0092)
Age				−0.0009***	−0.0010***	−0.0011***	−0.0009***
				(0.0003)	(0.0003)	(0.0003)	(0.0003)
Black				−0.0272***	−0.0307***	−0.0284***	−0.0390***
				(0.0096)	(0.0096)	(0.0096)	(0.0095)
Asian				−0.1301***	−0.1370***	−0.1285***	−0.1310***
				(0.0413)	(0.0413)	(0.0413)	(0.0408)
Indigenous American				0.0284	0.0098	0.0216	0.0172
				(0.0968)	(0.0969)	(0.0967)	(0.0957)
Skin-Tone				−0.0196	−0.0149	−0.0202	0.0083
				(0.0194)	(0.0194)	(0.0194)	(0.0193)
Attractiveness				−0.0096***	−0.0006	−0.0048	−0.0043
				(0.0034)	(0.0035)	(0.0035)	(0.0035)
Competence				−0.0079*	−0.0045	−0.0047	−0.0026
				(0.0042)	(0.0043)	(0.0043)	(0.0042)
Dominance				0.0045	0.0016	0.0050*	0.0032
				(0.0029)	(0.0029)	(0.0029)	(0.0029)
Trustworthiness				−0.0044	−0.0060	−0.0020	−0.0009
				(0.0041)	(0.0041)	(0.0041)	(0.0040)
Human Guess				0.0099	0.0096	0.0039	−0.0124
				(0.0214)	(0.0214)	(0.0214)	(0.0212)
Constant	0.3558***	0.3366***	0.4283***	0.3979***	0.3488***	0.4315***	0.2247***
	(0.0140)	(0.0129)	(0.0175)	(0.0313)	(0.0301)	(0.0321)	(0.0340)
Naive-AUC	0.551	0.544	0.565	0.589	0.58	0.593	0.632
Observations	14,250	14,250	14,250	14,250	14,250	14,250	14,250
Adjusted R^2	0.0055	0.0046	0.0087	0.0173	0.0151	0.0187	0.0383

Notes: This table replicates the analysis from Table VI but applies it to the in-time hold-out data, consisting of randomly selected cases arrested prior to July 17, 2019. The table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables. The algorithmic predictions of the judges' detain decision come from a convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. 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. 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 and unknown gender.

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

Table A.XXVIII: Is the algorithm rediscovering known facial features? (lock-box hold-out data, OOS by time)

	Algo Judge Detain Prediction				
	(1)	(2)	(3)	(4)	(5)
Male	0.0611***	0.0608***	0.0601***	0.0606***	0.0604***
	(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Age		−0.0001*	−0.0001**	−0.0001	−0.0001
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black		0.0144***	0.0101***	0.0099***	0.0096***
		(0.0018)	(0.0023)	(0.0023)	(0.0023)
Asian		0.0016	−0.0002	0.0004	0.0004
		(0.0086)	(0.0087)	(0.0086)	(0.0086)
Indigenous American		−0.0102	−0.0109	−0.0138	−0.0122
		(0.0189)	(0.0189)	(0.0188)	(0.0187)
Skin-Tone			−0.0133***	−0.0143***	−0.0146***
			(0.0046)	(0.0046)	(0.0046)
Attractiveness				0.0006	0.0006
				(0.0008)	(0.0008)
Competence				−0.0033***	−0.0033***
				(0.0010)	(0.0010)
Dominance				−0.0005	−0.0007
				(0.0007)	(0.0007)
Trustworthiness				−0.0026***	−0.0024***
				(0.0009)	(0.0009)
Human Guess					0.0183***
					(0.0048)
Constant	0.1296***	0.1250***	0.1360***	0.1606***	0.1512***
	(0.0016)	(0.0031)	(0.0049)	(0.0063)	(0.0068)
Observations	4,759	4,759	4,759	4,759	4,759
Adjusted R^2	0.1785	0.1916	0.1929	0.2026	0.2049

Notes: This table replicates the analysis from Table II but applies it to the out-of-time hold-out data. The table presents the results of regressing an algorithmic prediction of judge detention decisions against each of the different explanatory variables as listed in the rows, where each column represents a different regression specification. The algorithm was trained using mugshots from the training data set, and evaluated on the out-of-time partition of the hold-out set, including all valid arrests from the last 6 months of the data period (from July 17, 2019, to January 17, 2020). Data on skin tone, attractiveness, competence, dominance, and trustworthiness comes from asking subjects to assign feature ratings to mugshot images from the Mecklenburg County, NC Sheriff's Office public website. The human guess about the judges' decision comes from showing workers on the Prolific platform pairs of mugshot images and asking them to report which defendant they believe the judge would be more likely to detain. Regressions follow a linear probability model and also include indicators for unknown race and unknown gender.

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

Table A.XXIX: Does algorithm predict judge behavior after controlling for known factors? (lockbox hold-out data, OOS by time)

	Dependent variable:						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algo Judge Detain Prediction	0.9032***					0.8056***	0.8003***
	(0.0817)					(0.0915)	(0.0916)
Male		0.0788***	0.0791***		0.0776***	0.0288**	0.0288**
		(0.0119)	(0.0120)		(0.0124)	(0.0135)	(0.0135)
Age		0.0004	0.0004		0.0002	0.0003	0.0003
		(0.0004)	(0.0004)		(0.0005)	(0.0005)	(0.0005)
Black		−0.0044	−0.0026		−0.0025	−0.0105	−0.0110
		(0.0112)	(0.0146)		(0.0147)	(0.0146)	(0.0146)
Asian		−0.0396	−0.0388		−0.0395	−0.0398	−0.0397
		(0.0546)	(0.0548)		(0.0547)	(0.0543)	(0.0543)
Indigenous American		−0.0323	−0.0320		−0.0434	−0.0323	−0.0296
		(0.1194)	(0.1194)		(0.1192)	(0.1183)	(0.1183)
Skin-Tone			0.0054		0.0051	0.0166	0.0160
			(0.0290)		(0.0291)	(0.0289)	(0.0289)
Attractiveness				−0.0115**	−0.0065	−0.0070	−0.0069
				(0.0045)	(0.0050)	(0.0049)	(0.0049)
Competence				−0.0060	−0.0095	−0.0068	−0.0067
				(0.0061)	(0.0062)	(0.0062)	(0.0062)
Dominance				0.0048	−0.0001	0.0003	0.0001
				(0.0042)	(0.0043)	(0.0042)	(0.0042)
Trustworthiness				−0.0055	−0.0043	−0.0022	−0.0019
				(0.0057)	(0.0058)	(0.0057)	(0.0057)
Human Guess							0.0317
							(0.0302)
Constant	−0.0068	0.0824***	0.0779**	0.2157***	0.1689***	0.0395	0.0239
	(0.0152)	(0.0194)	(0.0308)	(0.0252)	(0.0400)	(0.0423)	(0.0448)
Naive-AUC	0.629	0.568	0.568	0.557	0.592	0.64	0.64
Observations	4,759	4,759	4,759	4,759	4,759	4,759	4,759
Adjusted R^2	0.0248	0.0089	0.0087	0.0044	0.0119	0.0276	0.0276

Notes: This table replicates the analysis from Table III but applies it to the out-of-time hold-out data. The table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables in the out-of-time partition of the hold-out set, including all valid arrests from the last 6 months of the data period (from July 17, 2019, to January 17, 2020). The algorithmic predictions of the judges' detain decision come from our convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. Measures of defendant demographics and current arrest charge come from government administrative data obtained from a combination of Mecklenburg County, NC and state agencies. Measures of skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXX: Correlation between well-groomed (first novel feature) and algorithm's prediction (lock-box hold-out data, OOS by time)

	Dependent variable:					
	Algo Judge Detain Prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Well-Groomed	−0.0067***	−0.0068***	−0.0075***	−0.0076***	−0.0073***	−0.0071***
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0008)	(0.0008)
Male		0.0611***	0.0610***	0.0602***	0.0610***	0.0609***
		(0.0019)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Age			−0.0003***	−0.0003***	−0.0002***	−0.0002***
			(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black			0.0140***	0.0094***	0.0091***	0.0089***
			(0.0017)	(0.0023)	(0.0023)	(0.0023)
Asian			0.0031	0.0012	0.0021	0.0020
			(0.0085)	(0.0085)	(0.0085)	(0.0085)
Indigenous American			−0.0167	−0.0175	−0.0174	−0.0160
			(0.0187)	(0.0186)	(0.0186)	(0.0186)
Skin-Tone				−0.0141***	−0.0150***	−0.0152***
				(0.0045)	(0.0045)	(0.0045)
Attractiveness					0.0025***	0.0025***
					(0.0008)	(0.0008)
Competence					−0.0017*	−0.0017*
					(0.0010)	(0.0010)
Dominance					−0.0002	−0.0004
					(0.0007)	(0.0007)
Trustworthiness					−0.0016*	−0.0015*
					(0.0009)	(0.0009)
Human Guess						0.0153***
						(0.0048)
Constant	0.2048***	0.1594***	0.1640***	0.1759***	0.1777***	0.1693***
	(0.0034)	(0.0033)	(0.0046)	(0.0060)	(0.0065)	(0.0070)
Observations	4,759	4,759	4,759	4,759	4,759	4,759
Adjusted R^2	0.0170	0.1959	0.2118	0.2133	0.2150	0.2166

Notes: This table replicates the analysis from Table IV but applies it to the out-of-time hold-out data, including all valid arrests from the last 6 months of the data period (from July 17, 2019, to January 17, 2020). The table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to hold-out set observations. Data on well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXXI: Correlation between heavy-faced (second novel feature) and algorithm's prediction (lock-box hold-out data, OOS by time)

	Dependent variable:						
	Algo Judge Detain Prediction						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	-0.0107^{***}	-0.0100^{***}	-0.0093^{***}	-0.0093^{***}	-0.0094^{***}	-0.0095^{***}	-0.0094^{***}
	(0.0007)	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Well-Groomed		-0.0052^{***}	-0.0053^{***}	-0.0061^{***}	-0.0061^{***}	-0.0057^{***}	-0.0056^{***}
		(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0008)	(0.0008)
Male			0.0601^{***}	0.0599^{***}	0.0589^{***}	0.0586^{***}	0.0585^{***}
			(0.0018)	(0.0018)	(0.0018)	(0.0019)	(0.0019)
Age				-0.0003^{***}	-0.0003^{***}	-0.0003^{***}	-0.0003^{***}
				(0.0001)	(0.0001)	(0.0001)	(0.0001)
Black				0.0143^{***}	0.0086^{***}	0.0084^{***}	0.0082^{***}
				(0.0017)	(0.0022)	(0.0022)	(0.0022)
Asian				0.0057	0.0034	0.0037	0.0036
				(0.0083)	(0.0083)	(0.0083)	(0.0083)
Indigenous American				−0.0136	−0.0146	−0.0143	−0.0132
				(0.0182)	(0.0182)	(0.0182)	(0.0182)
Skin-Tone					-0.0177^{***}	-0.0172^{***}	-0.0173^{***}
					(0.0044)	(0.0044)	(0.0044)
Attractiveness						0.0001	0.0001
						(0.0008)	(0.0008)
Competence						−0.0016	−0.0015
						(0.0010)	(0.0010)
Dominance						0.0011^{*}	0.0010
						(0.0007)	(0.0007)
Trustworthiness						0.0004	0.0004
						(0.0009)	(0.0009)
Human Guess							0.0122^{***}
							(0.0046)
Constant	0.2306^{***}	0.2498^{***}	0.2019^{***}	0.2061^{***}	0.2216^{***}	0.2194^{***}	0.2124^{***}
	(0.0036)	(0.0044)	(0.0043)	(0.0053)	(0.0065)	(0.0069)	(0.0074)
Observations	4,759	4,759	4,759	4,759	4,759	4,759	4,759
Adjusted R^2	0.0515	0.0612	0.2338	0.2497	0.2520	0.2522	0.2531

Notes: This table replicates the analysis from Table V but applies it to the out-of-time hold-out data, including all valid arrests from the last 6 months of the data period (from July 17, 2019, to January 17, 2020). The table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables. Algorithmic predictions of judges' decisions come from applying an algorithm built with face images in the training data set to hold-out set observations. Data on heavy-faced, well-groomed, skin tone, attractiveness, competence, dominance, and trustworthiness come from subject ratings of mugshot images. 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 and unknown gender.

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

Table A.XXXII: Do well-groomed and heavy-faced (first and second novel features) correlate with judge decisions? (lock-box hold-out data, OOS by time)

	Dependent variable:						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	-0.0256^{***}		-0.0230^{***}	-0.0241^{***}		-0.0227^{***}	-0.0162^{***}
	(0.0039)		(0.0039)	(0.0040)		(0.0040)	(0.0041)
Well-Groomed		-0.0233^{***}	-0.0198^{***}		-0.0187^{***}	-0.0150^{***}	-0.0111^{**}
		(0.0042)	(0.0043)		(0.0054)	(0.0054)	(0.0054)
Algo Judge Detain Prediction							0.6922^{***}
							(0.0944)
Male				0.0712^{***}	0.0785^{***}	0.0727^{***}	0.0322^{**}
				(0.0124)	(0.0124)	(0.0124)	(0.0135)
Age				−0.00004	−0.00003	−0.0002	−0.00004
				(0.0005)	(0.0005)	(0.0005)	(0.0005)
Black				−0.0055	−0.0051	−0.0069	−0.0125
				(0.0146)	(0.0147)	(0.0146)	(0.0145)
Asian				−0.0345	−0.0351	−0.0312	−0.0337
				(0.0545)	(0.0546)	(0.0545)	(0.0542)
Indigenous American				−0.0343	−0.0494	−0.0427	−0.0335
				(0.1188)	(0.1191)	(0.1188)	(0.1181)
Skin-Tone				−0.0013	0.0026	−0.0024	0.0096
				(0.0290)	(0.0291)	(0.0290)	(0.0289)
Attractiveness				-0.0115^{**}	−0.0016	−0.0073	−0.0074
				(0.0050)	(0.0051)	(0.0052)	(0.0052)
Competence				−0.0080	−0.0053	−0.0048	−0.0038
				(0.0062)	(0.0063)	(0.0063)	(0.0063)
Dominance				0.0033	0.0003	0.0037	0.0030
				(0.0043)	(0.0043)	(0.0043)	(0.0043)
Trustworthiness				0.0016	−0.0015	0.0033	0.0030
				(0.0058)	(0.0058)	(0.0059)	(0.0058)
Human Guess				0.0367	0.0386	0.0310	0.0226
				(0.0303)	(0.0304)	(0.0304)	(0.0302)
Constant	0.2837^{***}	0.2541^{***}	0.3569^{***}	0.2649^{***}	0.1926^{***}	0.2965^{***}	0.1495^{***}
	(0.0207)	(0.0193)	(0.0260)	(0.0473)	(0.0451)	(0.0486)	(0.0523)
Naive-AUC	0.575	0.563	0.588	0.618	0.603	0.622	0.651
Observations	4,759	4,759	4,759	4,759	4,759	4,759	4,759
Adjusted R^2	0.0088	0.0062	0.0131	0.0194	0.0145	0.0208	0.0316

Notes: This table replicates the analysis from Table VI but applies it to the out-of-time hold-out data, including all valid arrests from the last 6 months of the data period (from July 17, 2019, to January 17, 2020). The table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables. The algorithmic predictions of the judges' detain decision come from a convolutional neural network algorithm built using the defendants' face image as the only feature, using data from the training data set. 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. 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 and unknown gender.

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

Appendix Figures

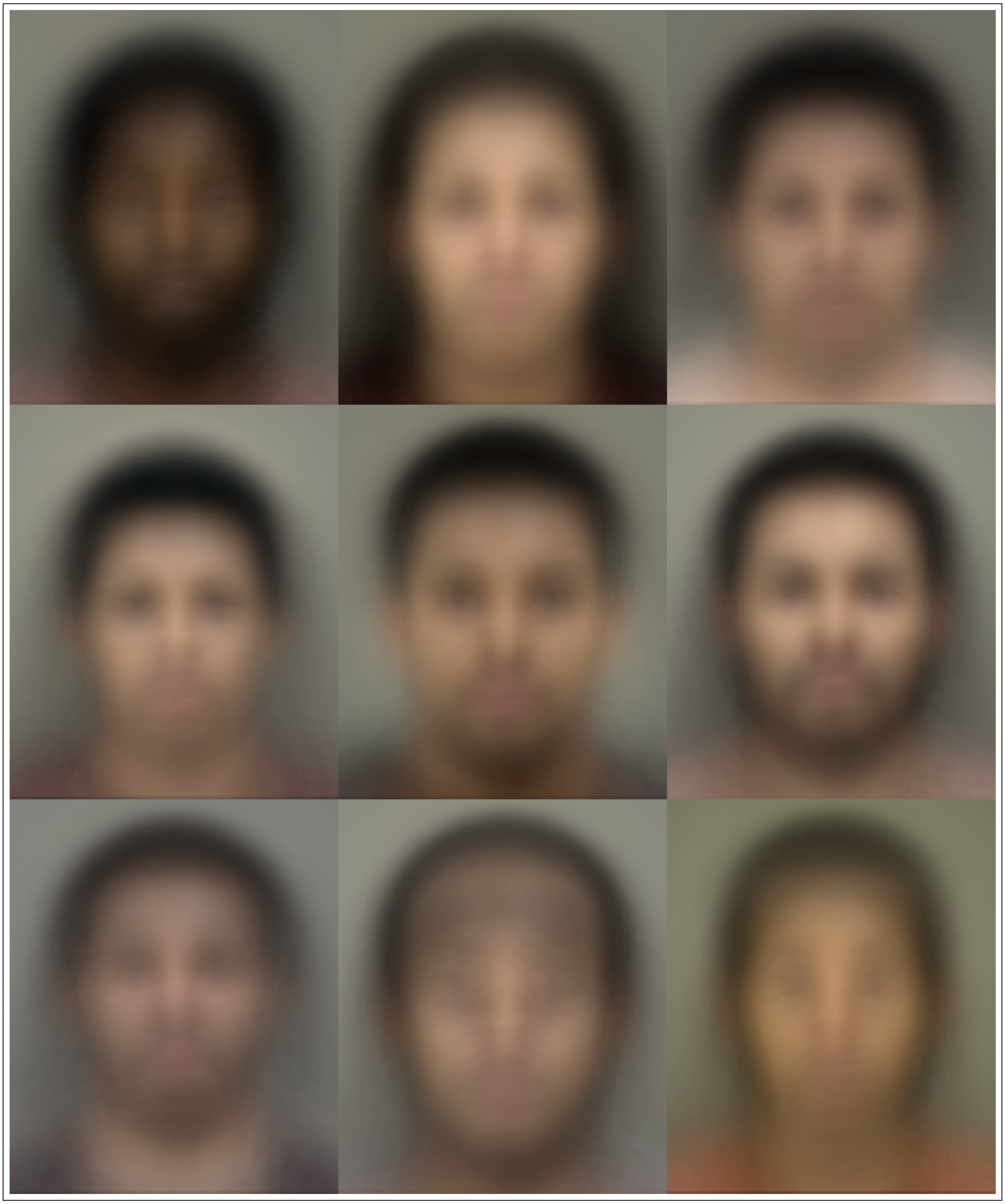


Figure A.I: Eigenfaces

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

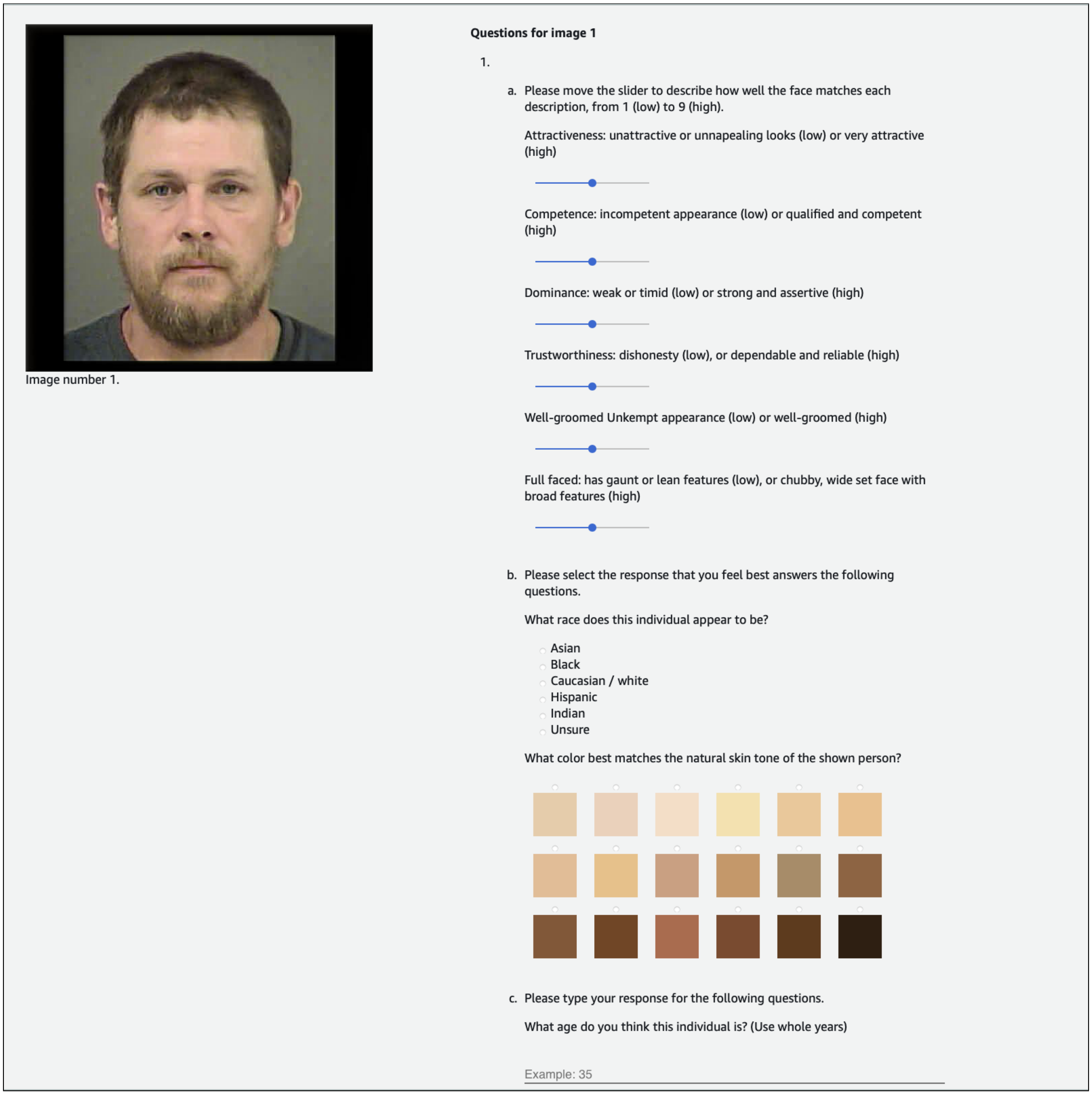


Image number 1.

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.

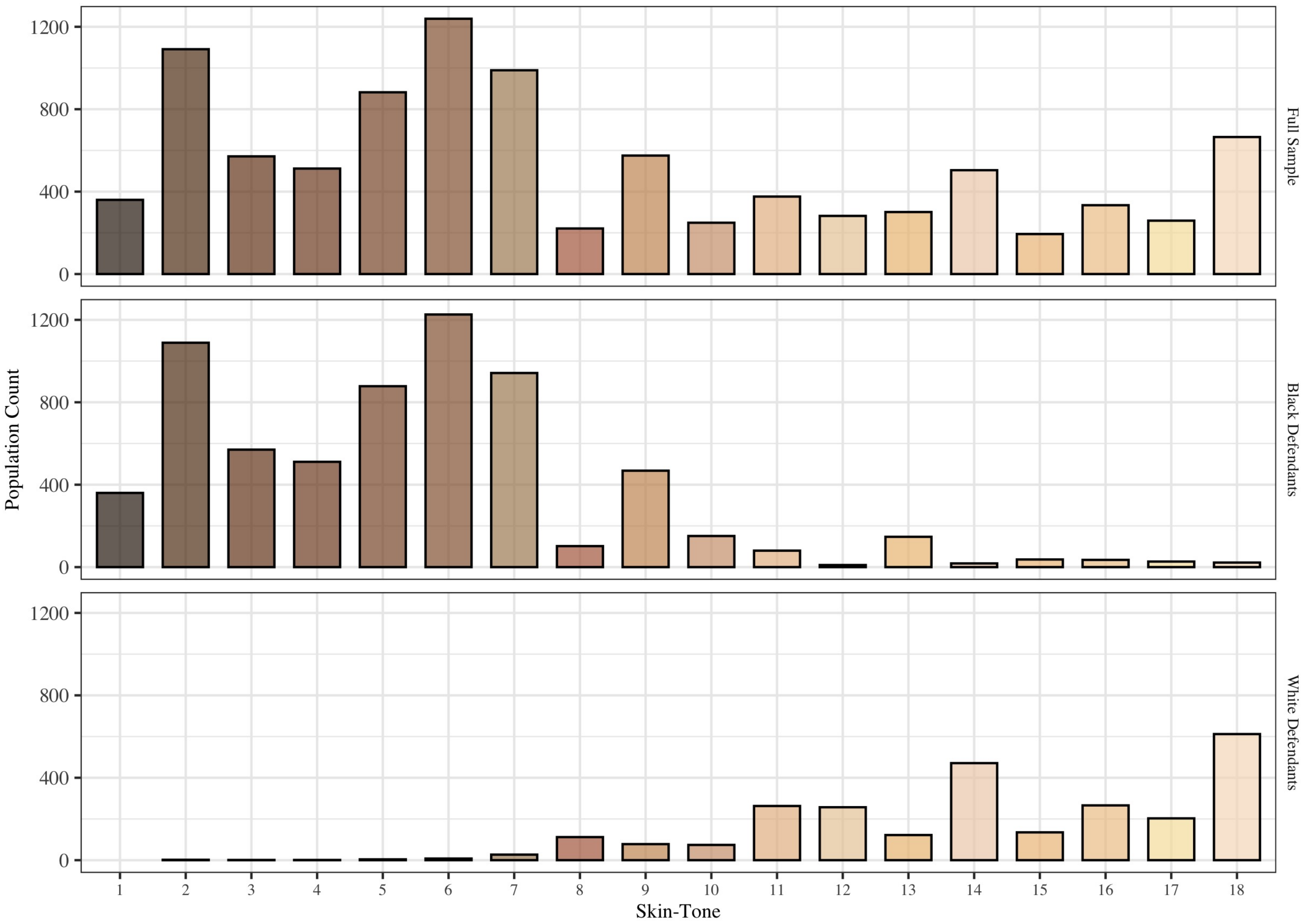
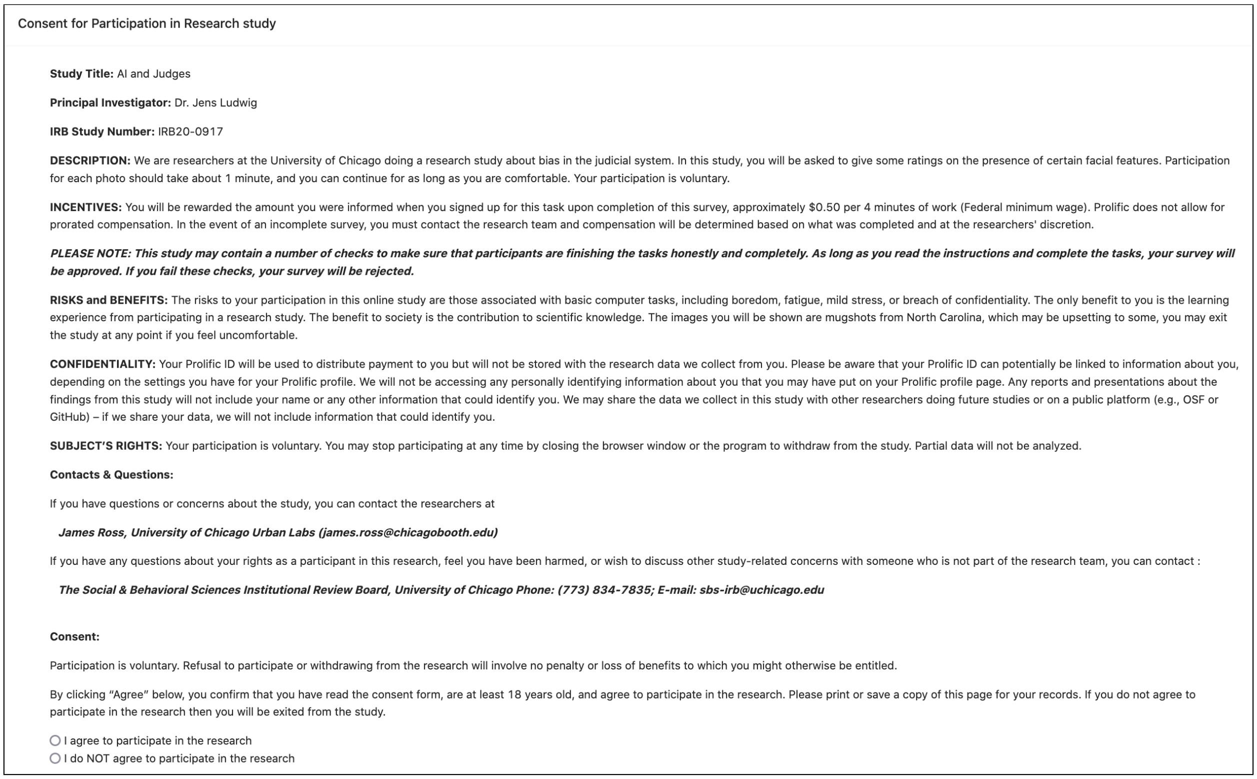


Figure A.III: Distribution of skin-tone categories for full validation sample, and by defendant race

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

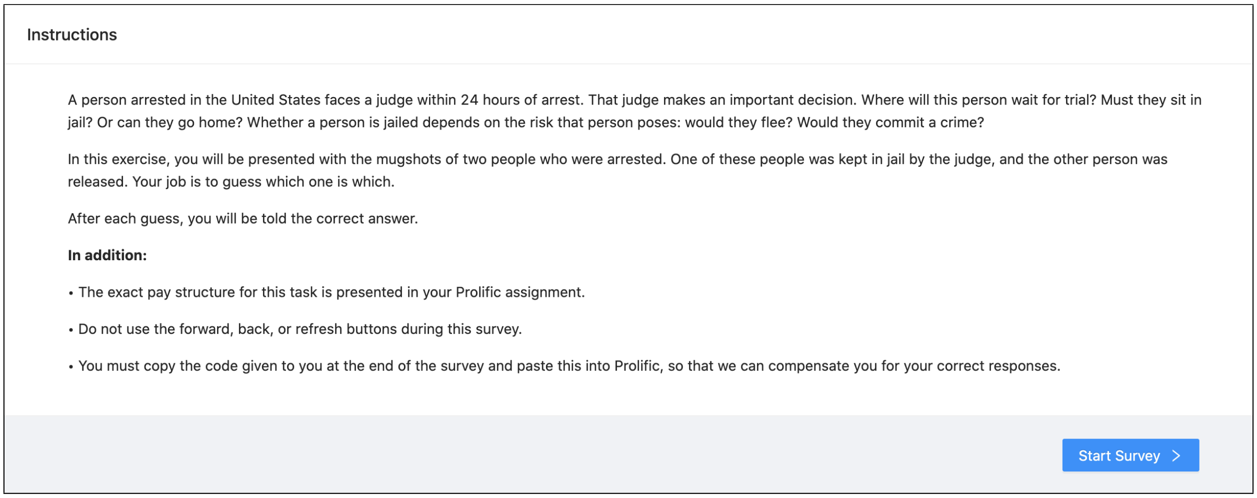
Consent:

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:

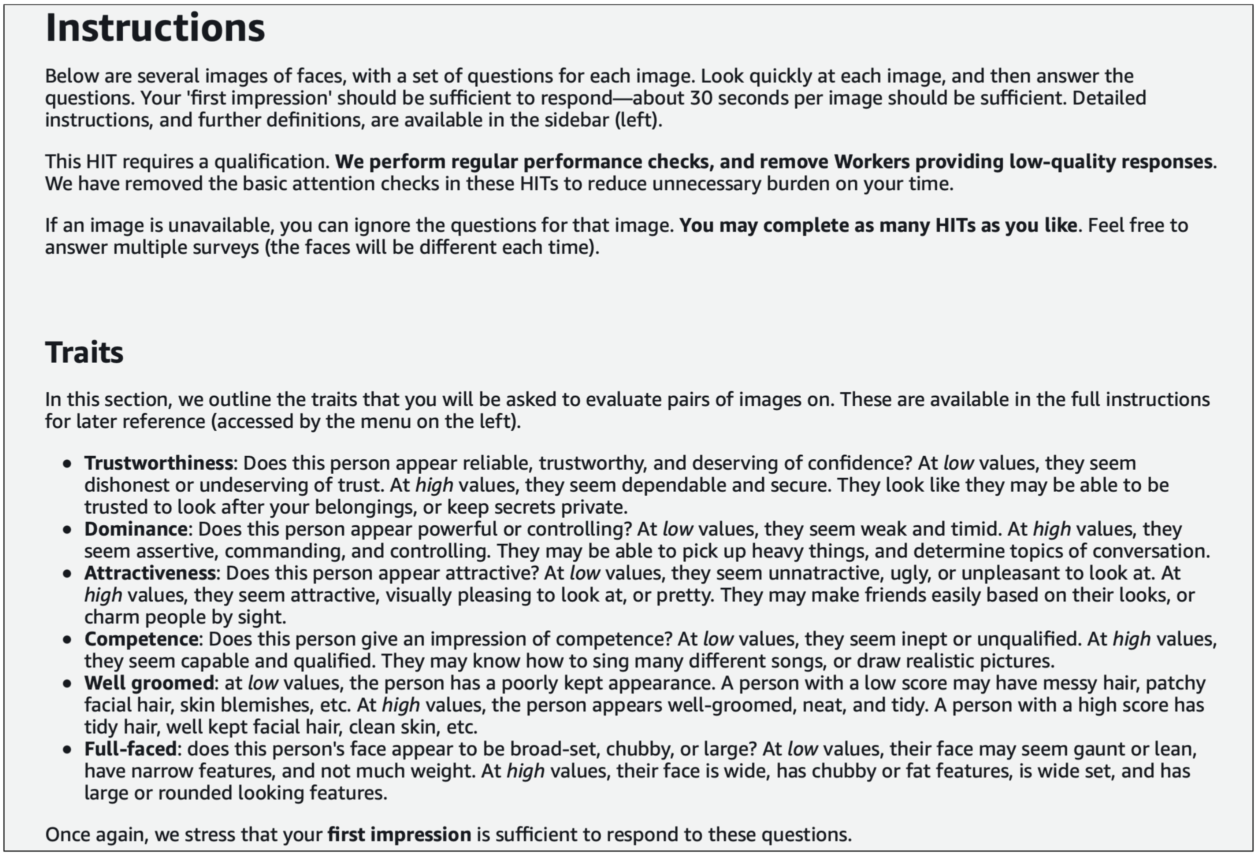
• The exact pay structure for this task is presented in your Prolific assignment.

• 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.

(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

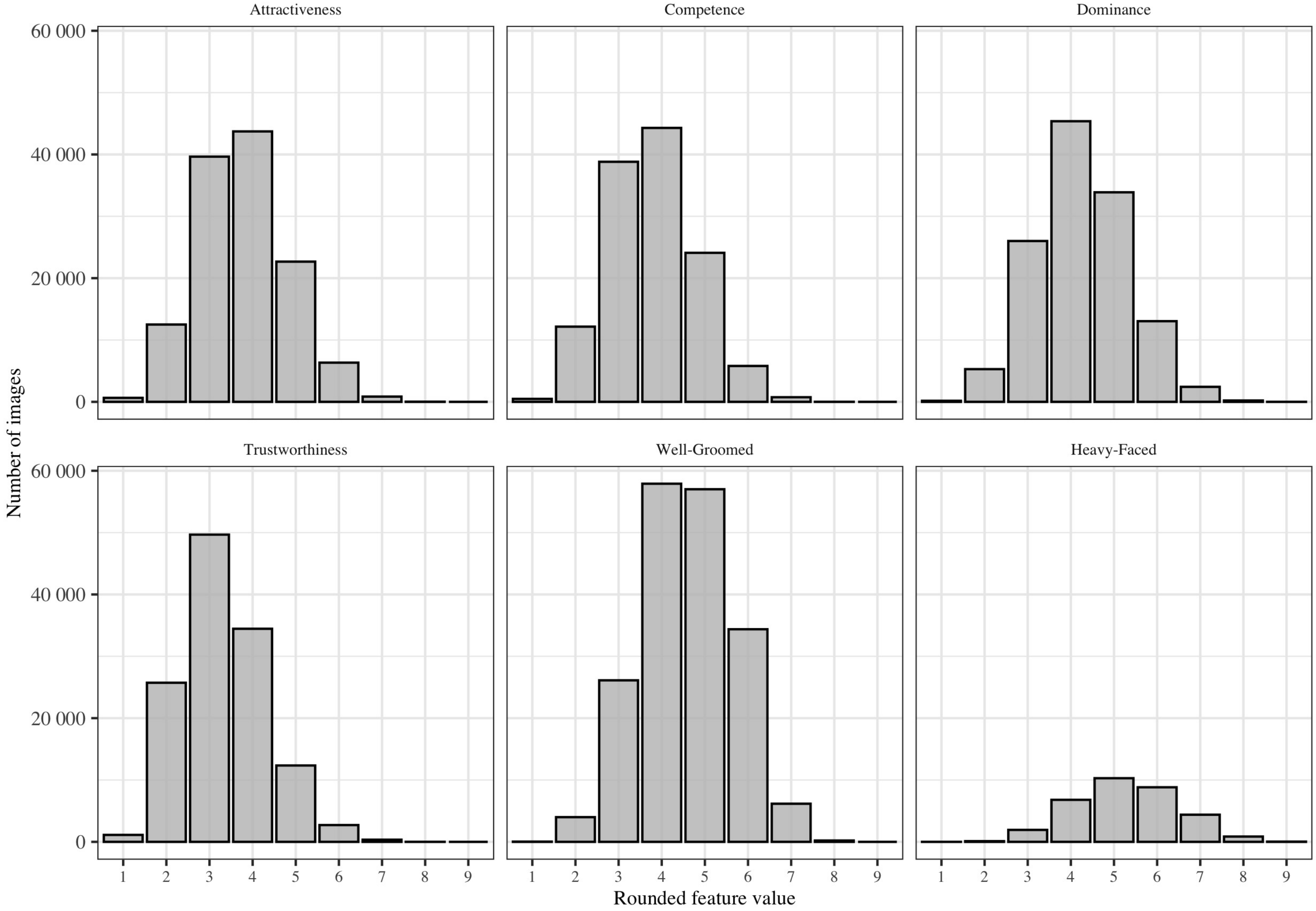


Figure A.VI: Distribution of human ratings of psychological features based on face images

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).

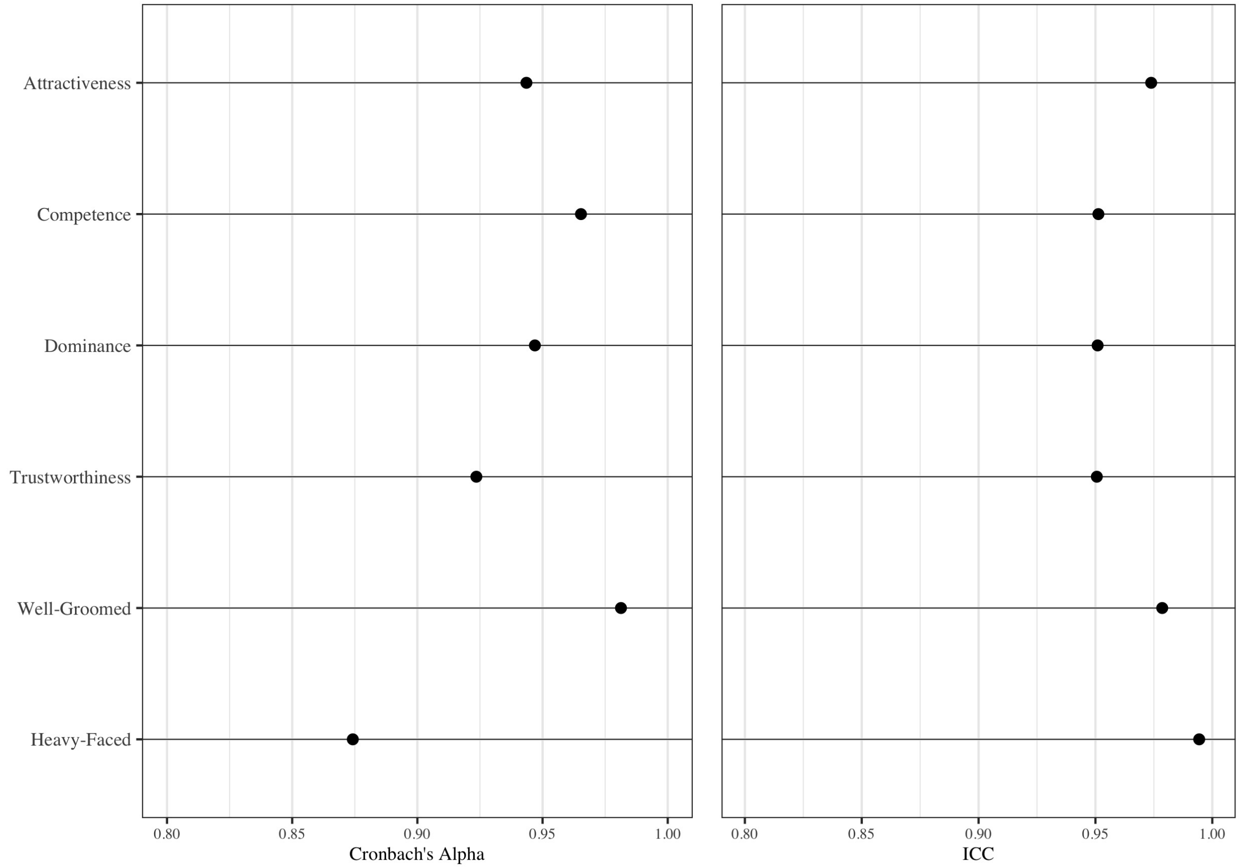


Figure A.VII: Reliability measures for human-rated psychological features

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.

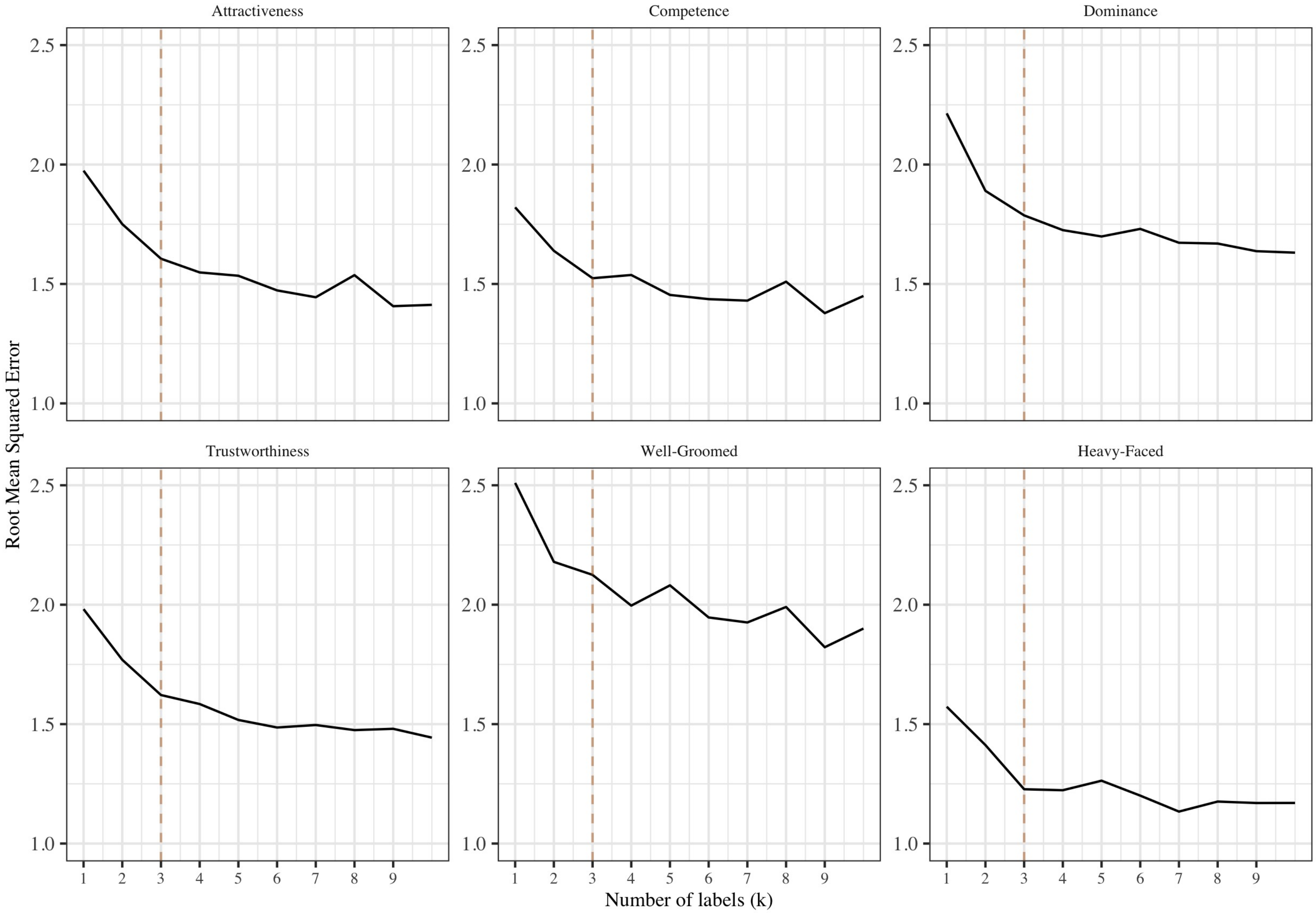
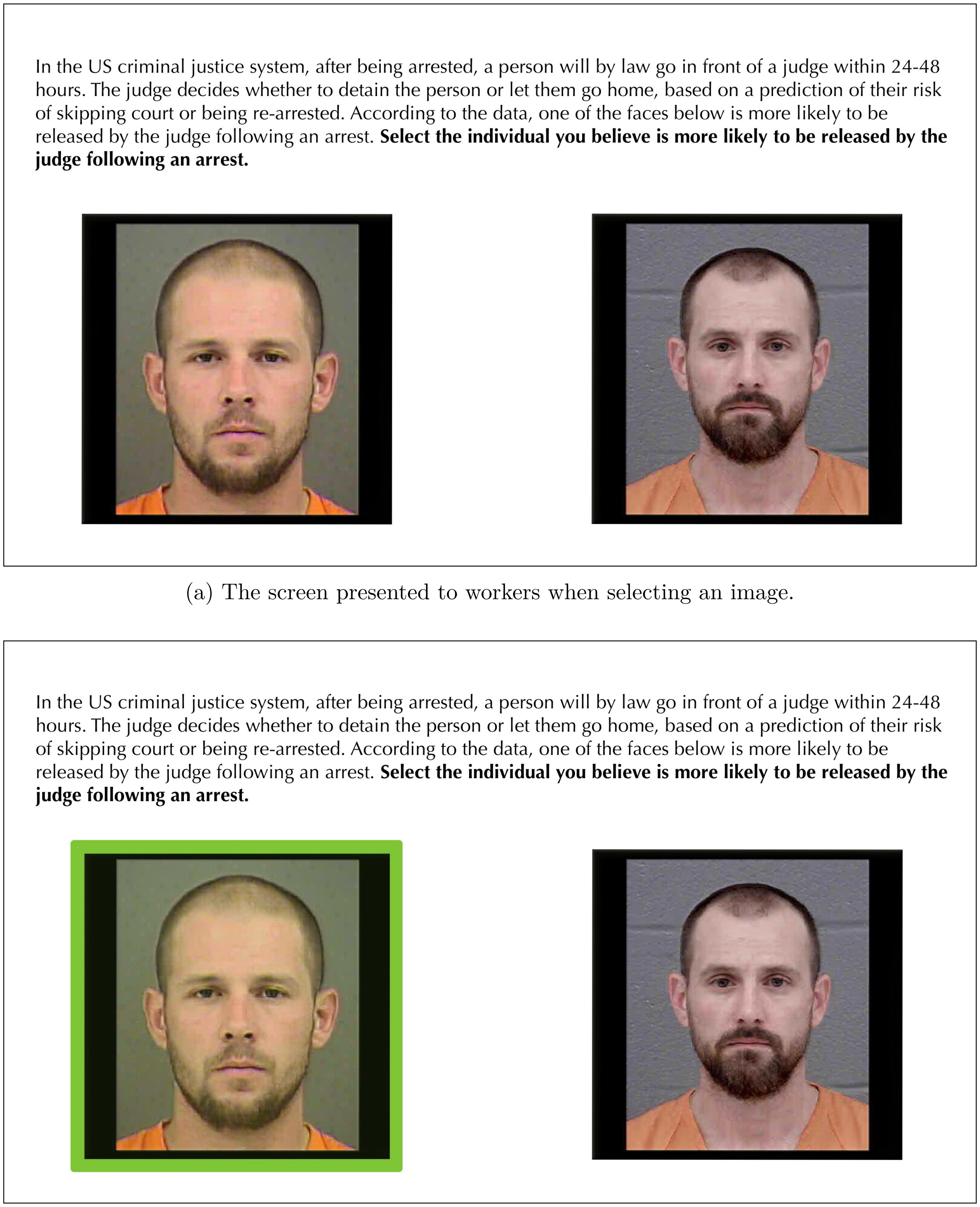


Figure A.VIII: Signal vs. noise in human ratings by number of ratings provided

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.



In the US criminal justice system, after being arrested, a person will by law go in front of a judge within 24-48 hours. The judge decides whether to detain the person or let them go home, based on a prediction of their risk of skipping court or being re-arrested. According to the data, one of the faces below is more likely to be released by the judge following an arrest. **Select the individual you believe is more likely to be released by the judge following an arrest.**

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

In the US criminal justice system, after being arrested, a person will by law go in front of a judge within 24-48 hours. The judge decides whether to detain the person or let them go home, based on a prediction of their risk of skipping court or being re-arrested. According to the data, one of the faces below is more likely to be released by the judge following an arrest. **Select the individual you believe is more likely to be released by the judge following an arrest.**

(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.

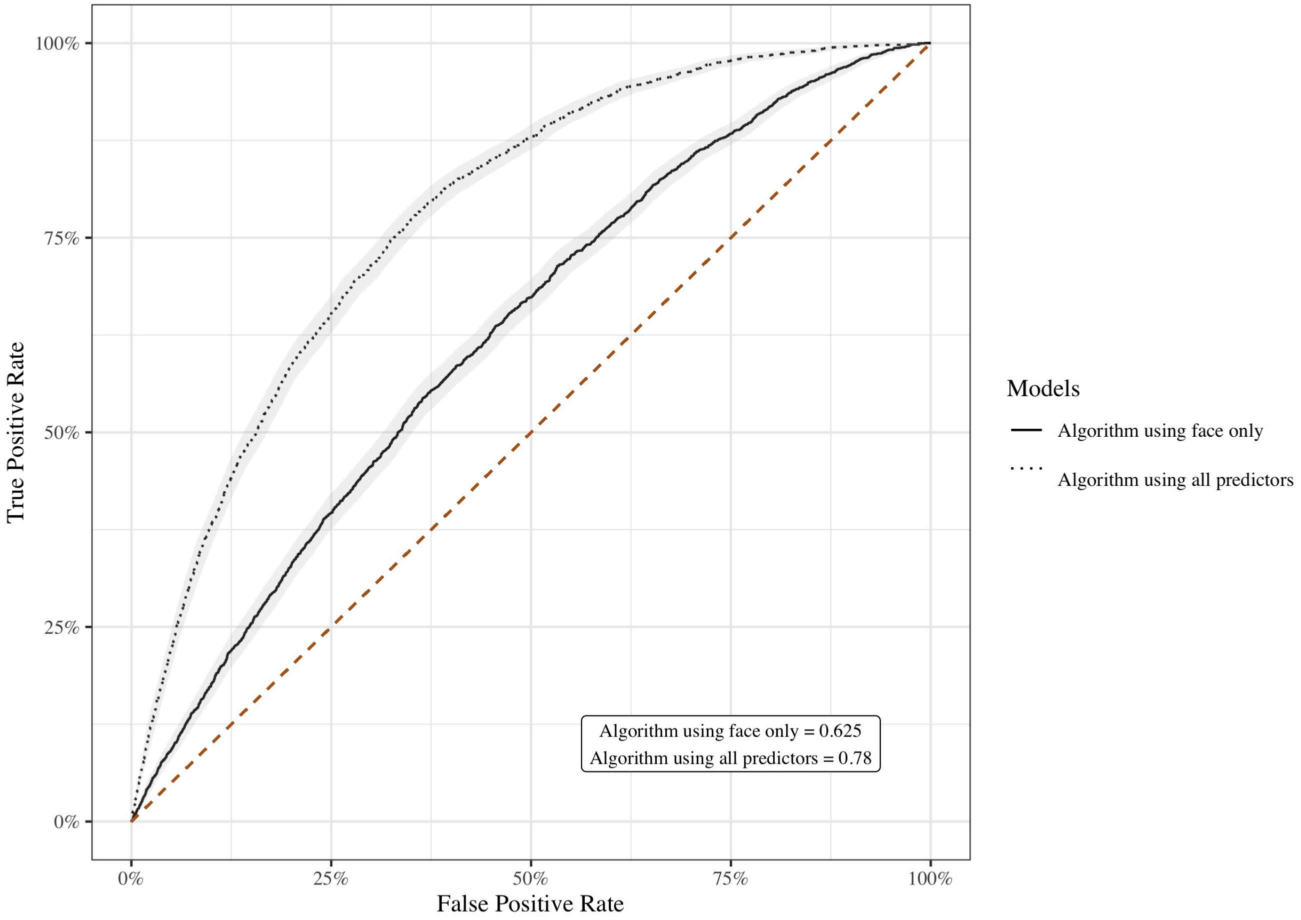


Figure A.X: Accuracy of algorithmic models of judge decisions

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.

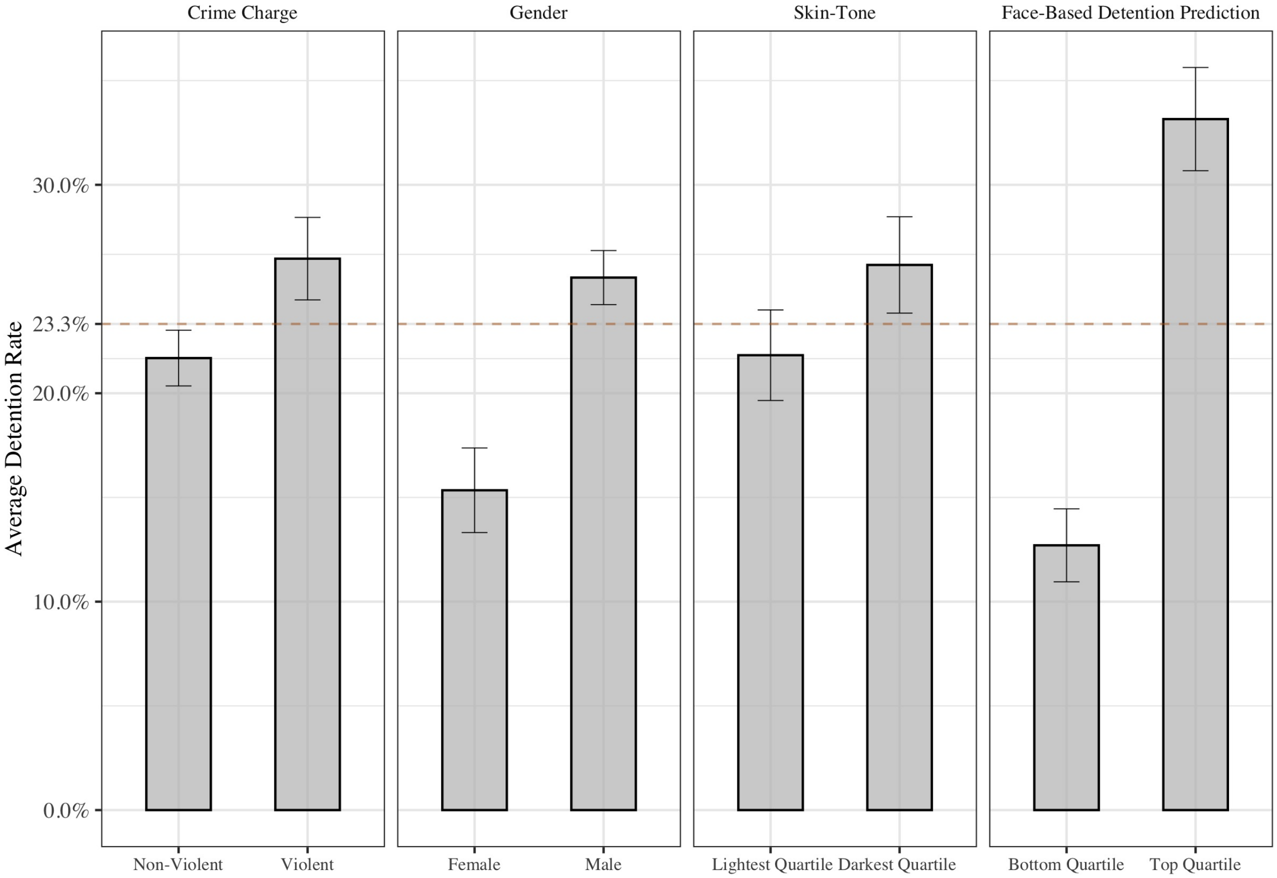
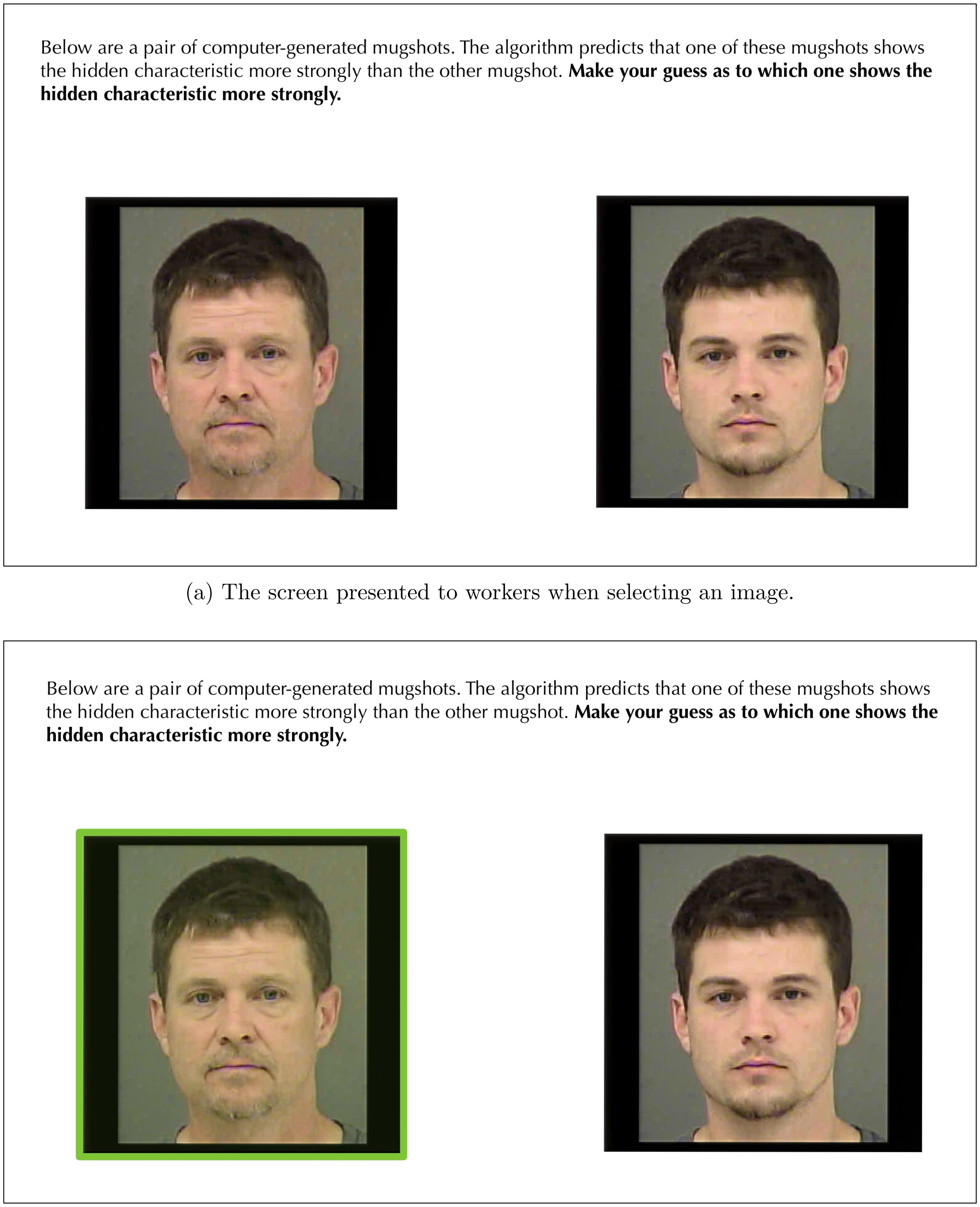


Figure A.XI: Relationship between detention rates and defendant characteristics

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%.



Below are a pair of computer-generated mugshots. The algorithm predicts that one of these mugshots shows the hidden characteristic more strongly than the other mugshot. **Make your guess as to which one shows the hidden characteristic more strongly.**

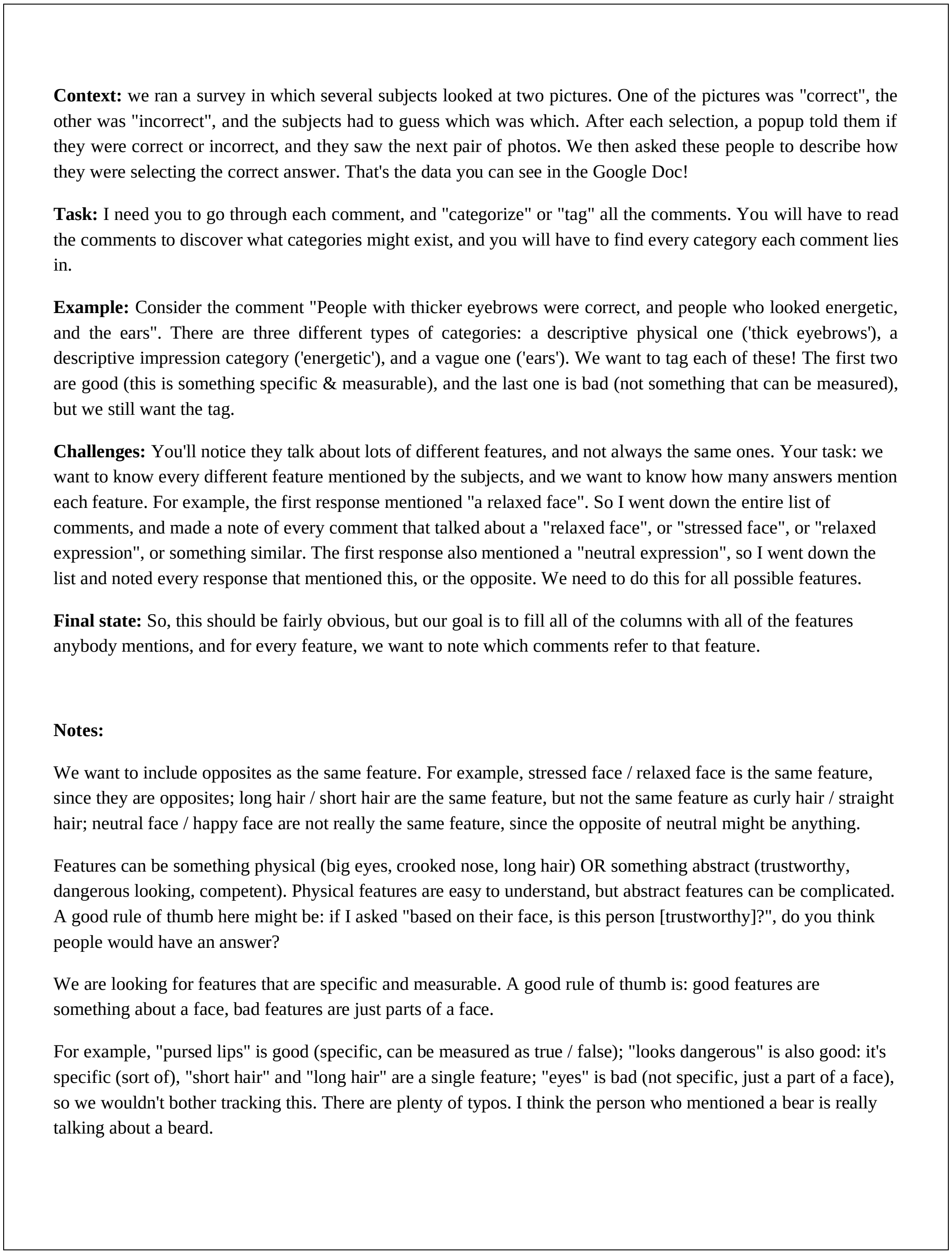
(a) The screen presented to workers when selecting an image.

Below are a pair of computer-generated mugshots. The algorithm predicts that one of these mugshots shows the hidden characteristic more strongly than the other mugshot. **Make your guess as to which one shows the hidden characteristic more strongly.**

(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:

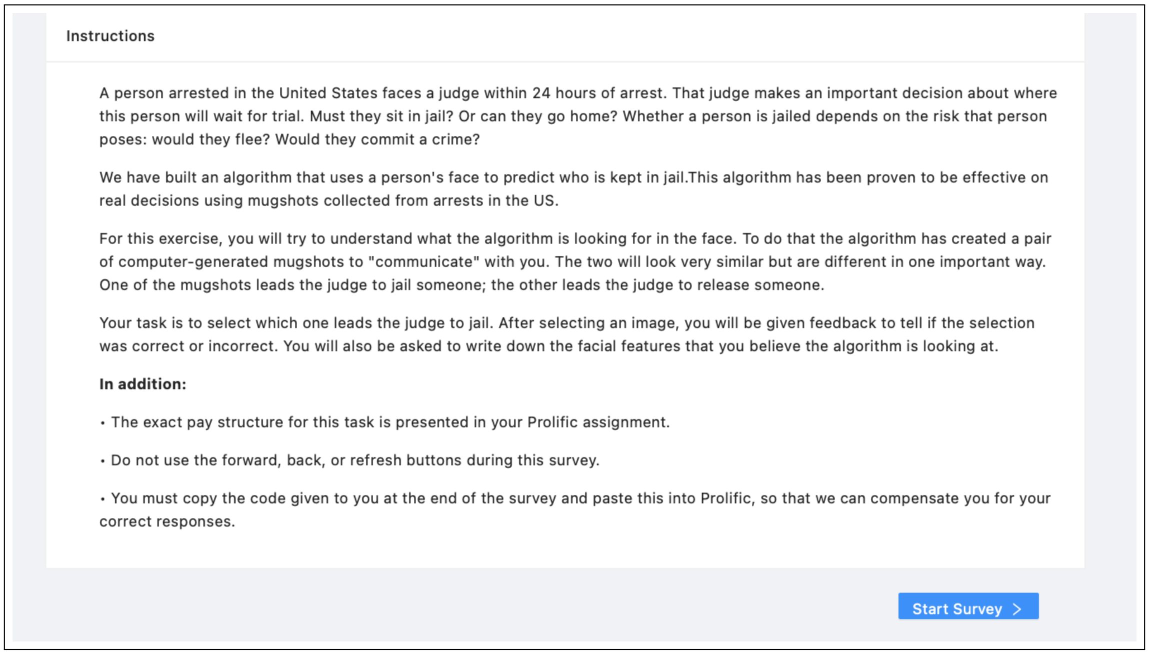
We want to include opposites as the same feature. For example, stressed face / relaxed face is the same feature, since they are opposites; long hair / short hair are the same feature, but not the same feature as curly hair / straight hair; neutral face / happy face are not really the same feature, since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



Instructions

A person arrested in the United States faces a judge within 24 hours of arrest. That judge makes an important decision about where this person will 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?

We have built an algorithm that uses a person's face to predict who is kept in jail.This algorithm has been proven to be effective on real decisions using mugshots collected from arrests in the US.

For this exercise, you will try to understand what the algorithm is looking for in the face. To do that the algorithm has created a pair of computer-generated mugshots to "communicate" with you. The two will look very similar but are different in one important way. One of the mugshots leads the judge to jail someone; the other leads the judge to release someone.

Your task is to select which one leads the judge to jail. After selecting an image, you will be given feedback to tell if the selection was correct or incorrect. You will also be asked to write down the facial features that you believe the algorithm is looking at.

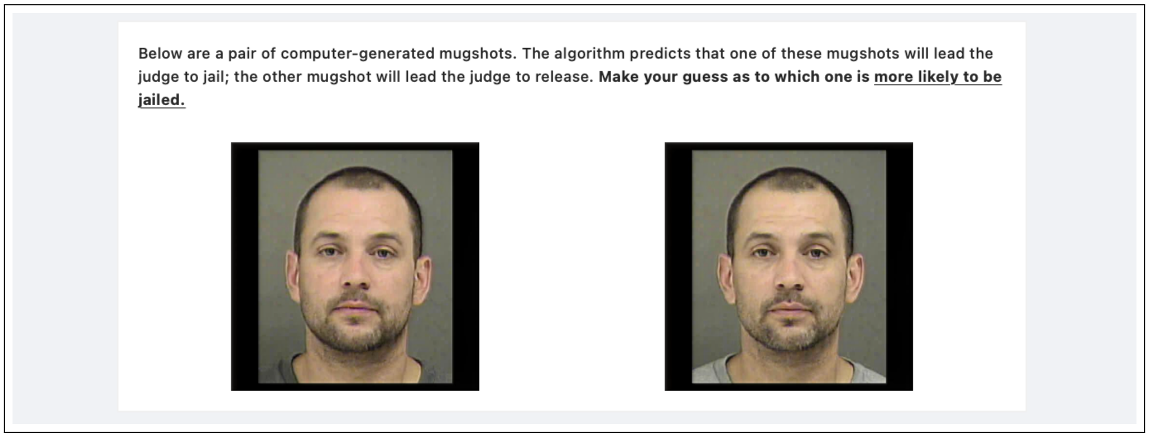
In addition:

• 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.

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



Below are a pair of computer-generated mugshots. The algorithm predicts that one of these mugshots will lead the judge to jail; the other mugshot will lead the judge to release. **Make your guess as to which one is <u>more likely to be jailed.</u>**

(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.

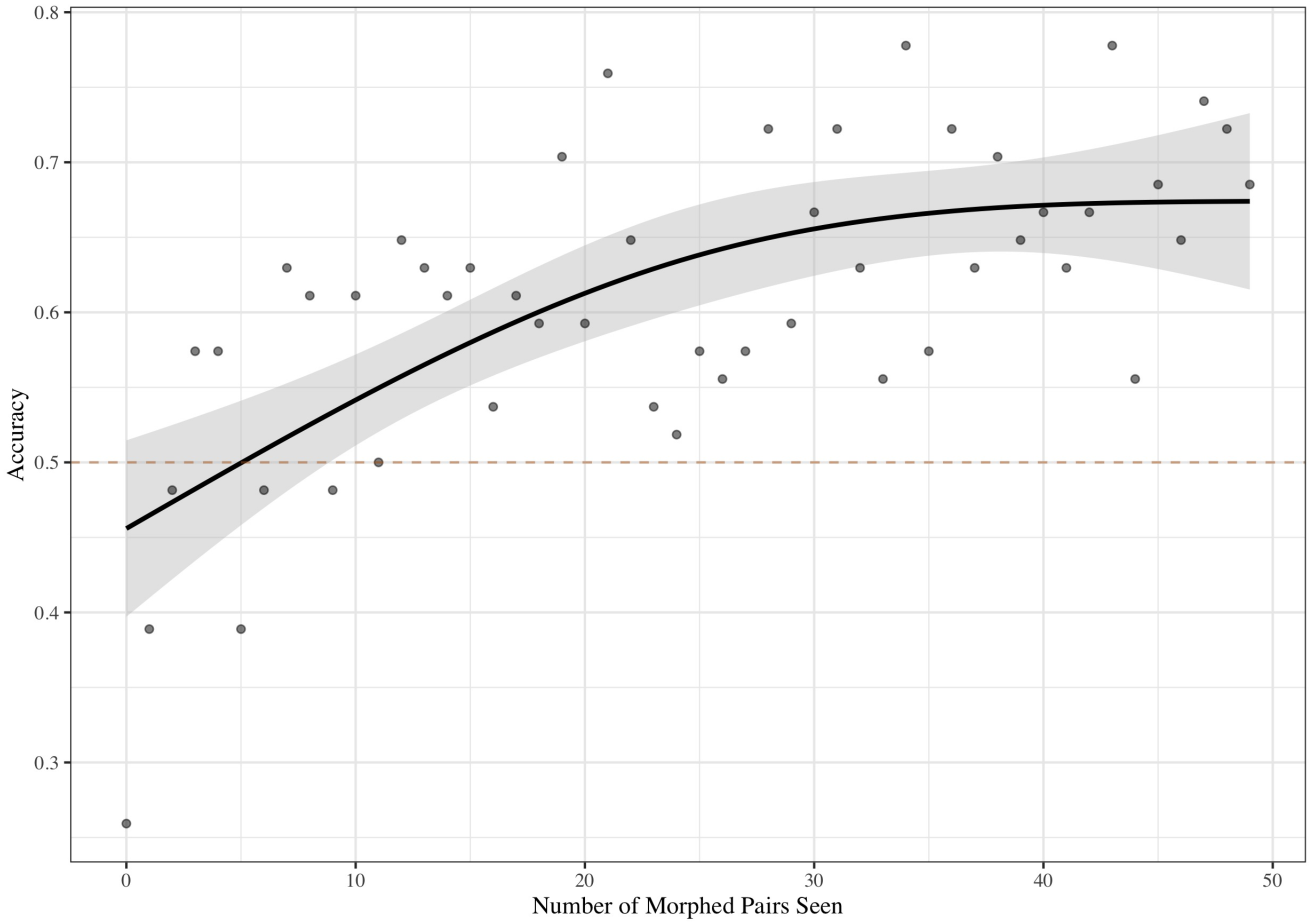


Figure A.XV: Subject performance guessing relative detention risk across morphed image pairs as a function of number of images seen

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.

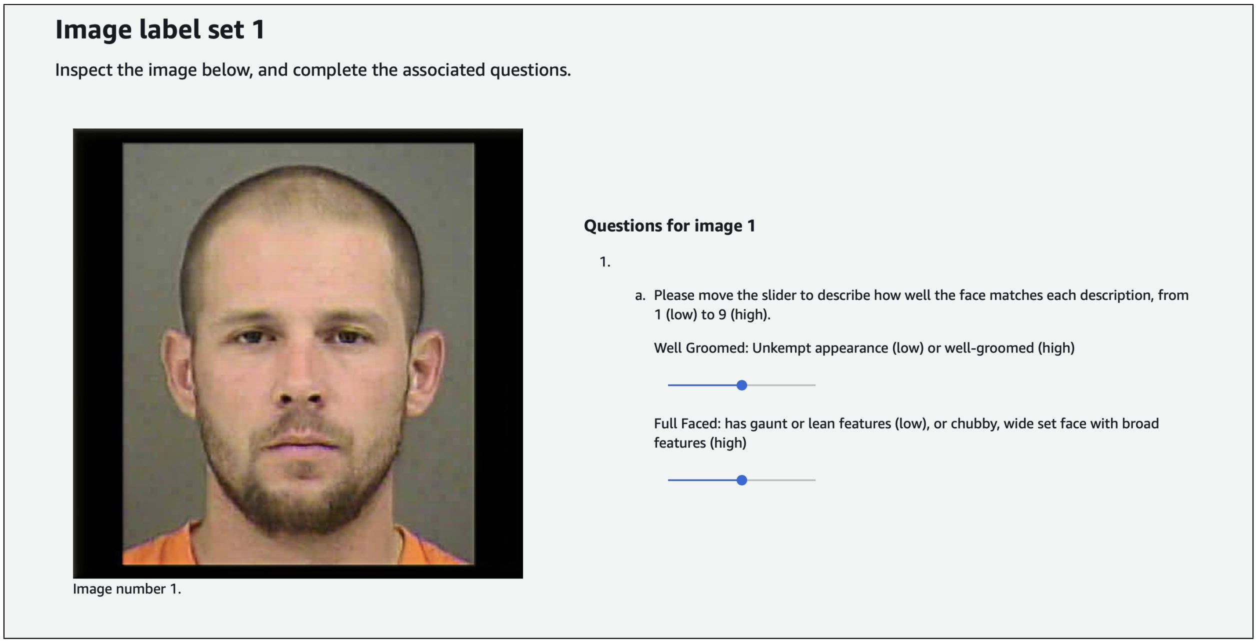


Image label set 1

Inspect the image below, and complete the associated questions.

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).

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)

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.

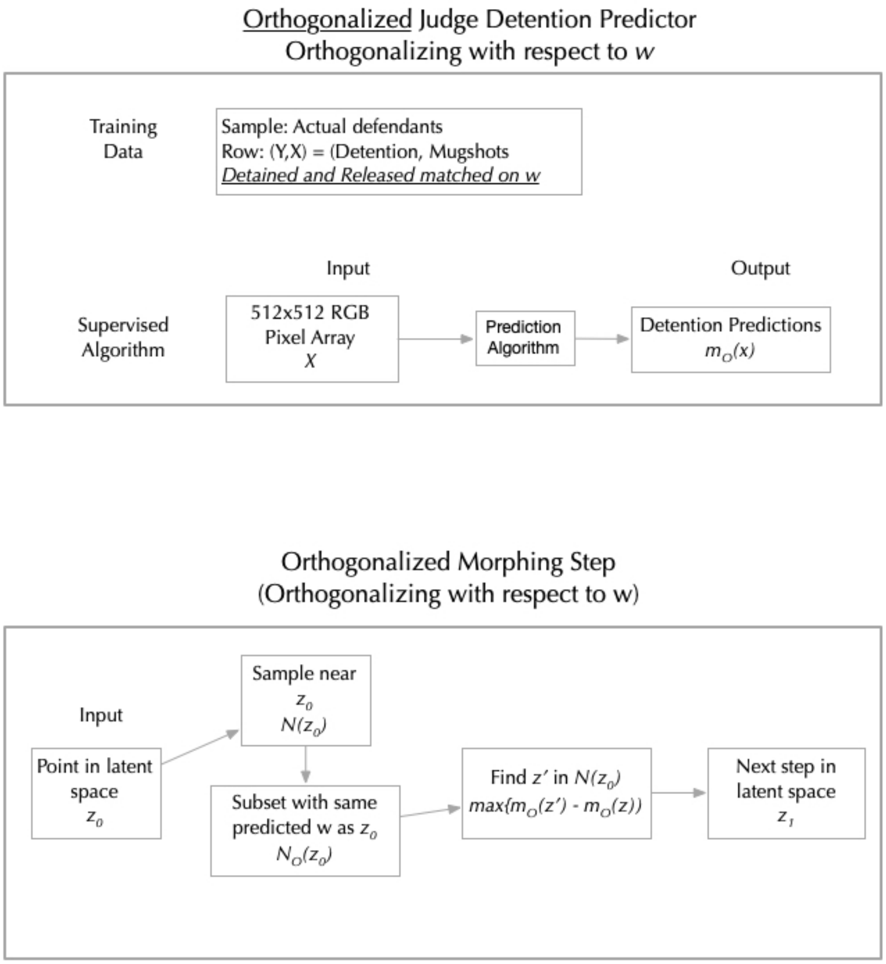
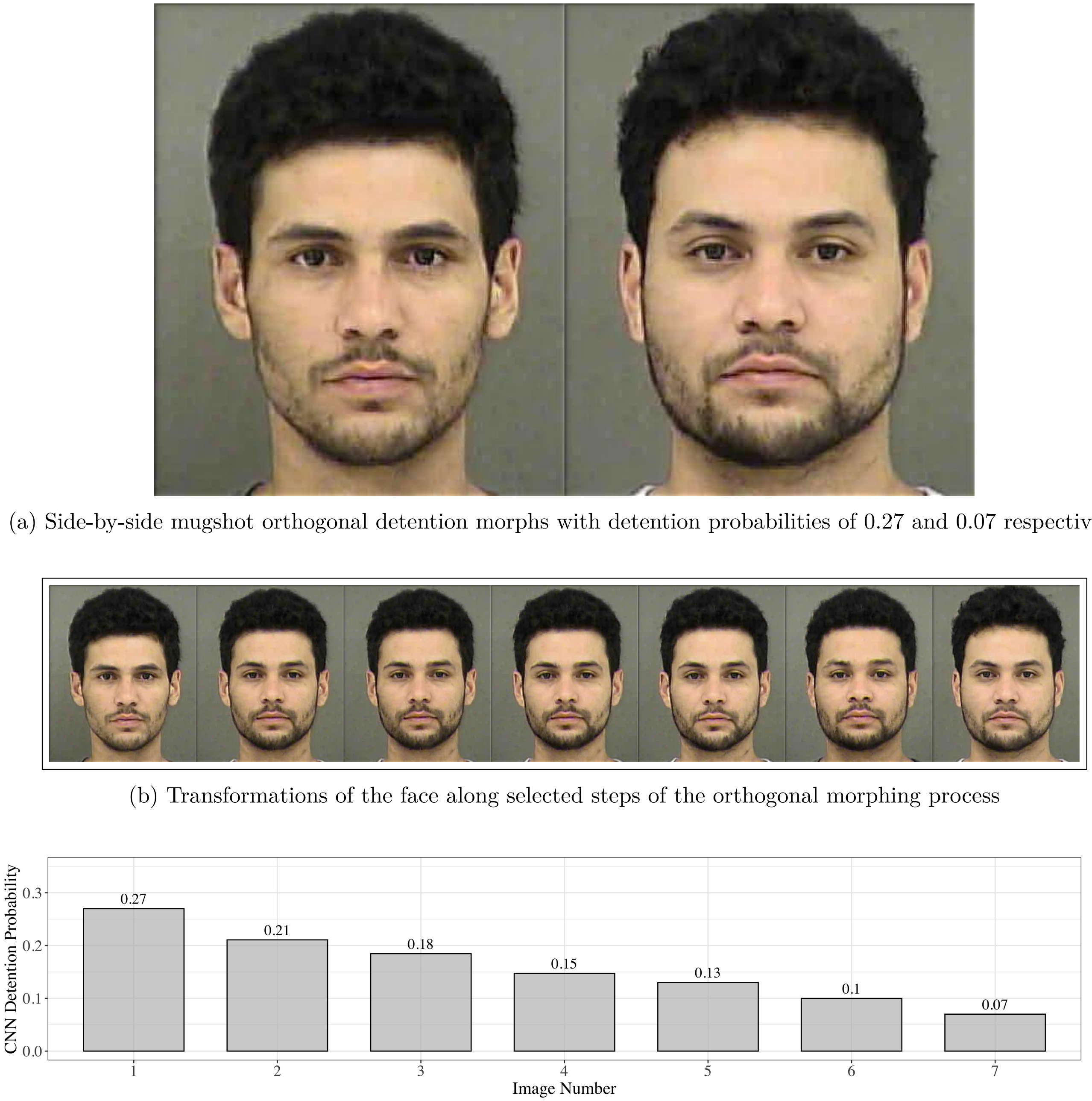

Orthogonalized Judge Detention Predictor
Orthogonalizing with respect to w
Training Data
Sample: Actual defendants
Row: (Y,X) = (Detention, Mugshots
Detained and Released matched on w
Input
Output
Supervised Algorithm
512x512 RGB Pixel Array X
Prediction Algorithm
Detention Predictions $m_O(x)$
Orthogonalized Morphing Step
(Orthogonalizing with respect to w)
Input
Point in latent space z_0
Sample near z_0 $N(z_0)$
Subset with same predicted w as z_0 $N_O(z_0)$
Find z' in $N(z_0)$ $max(m_O(z') - m_O(z))$
Next step in latent space z_1

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.

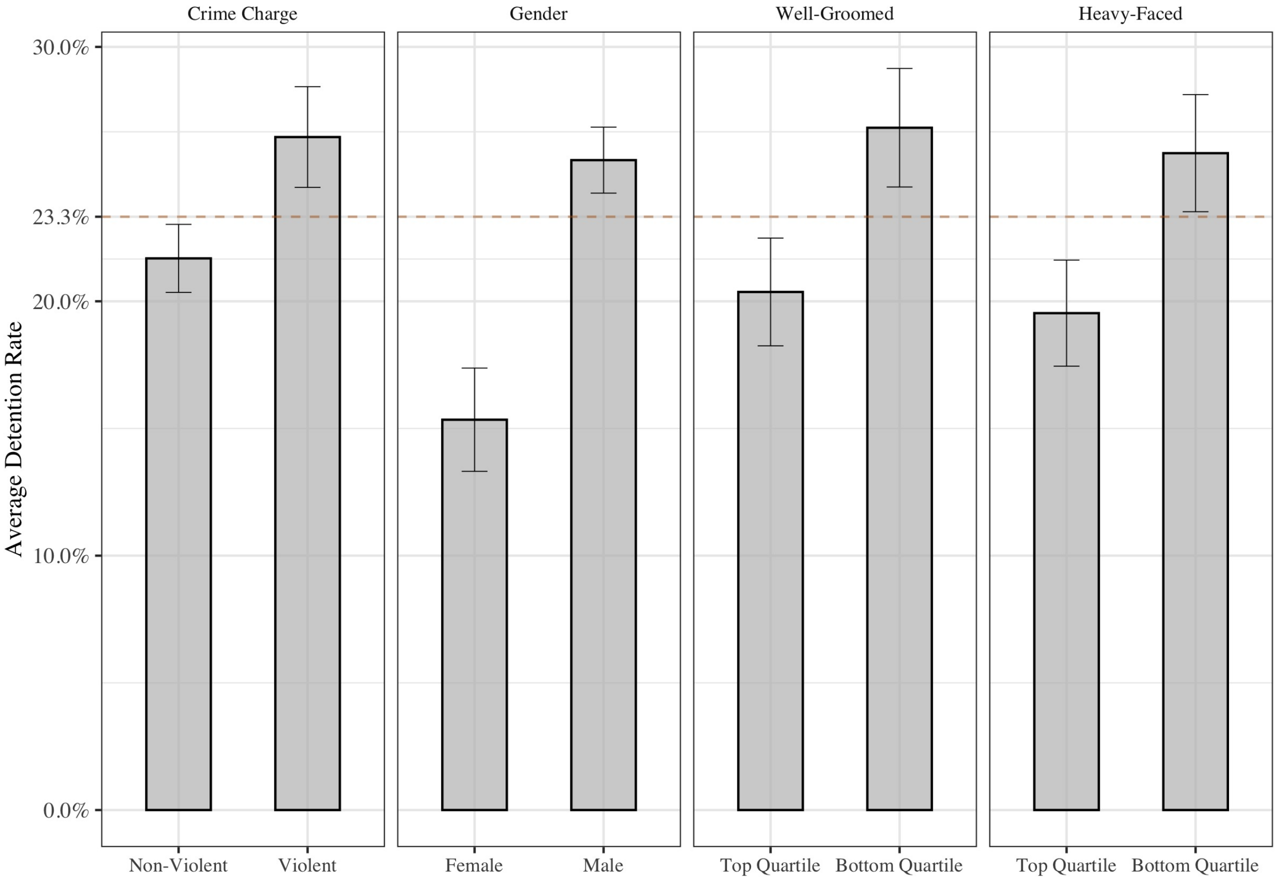
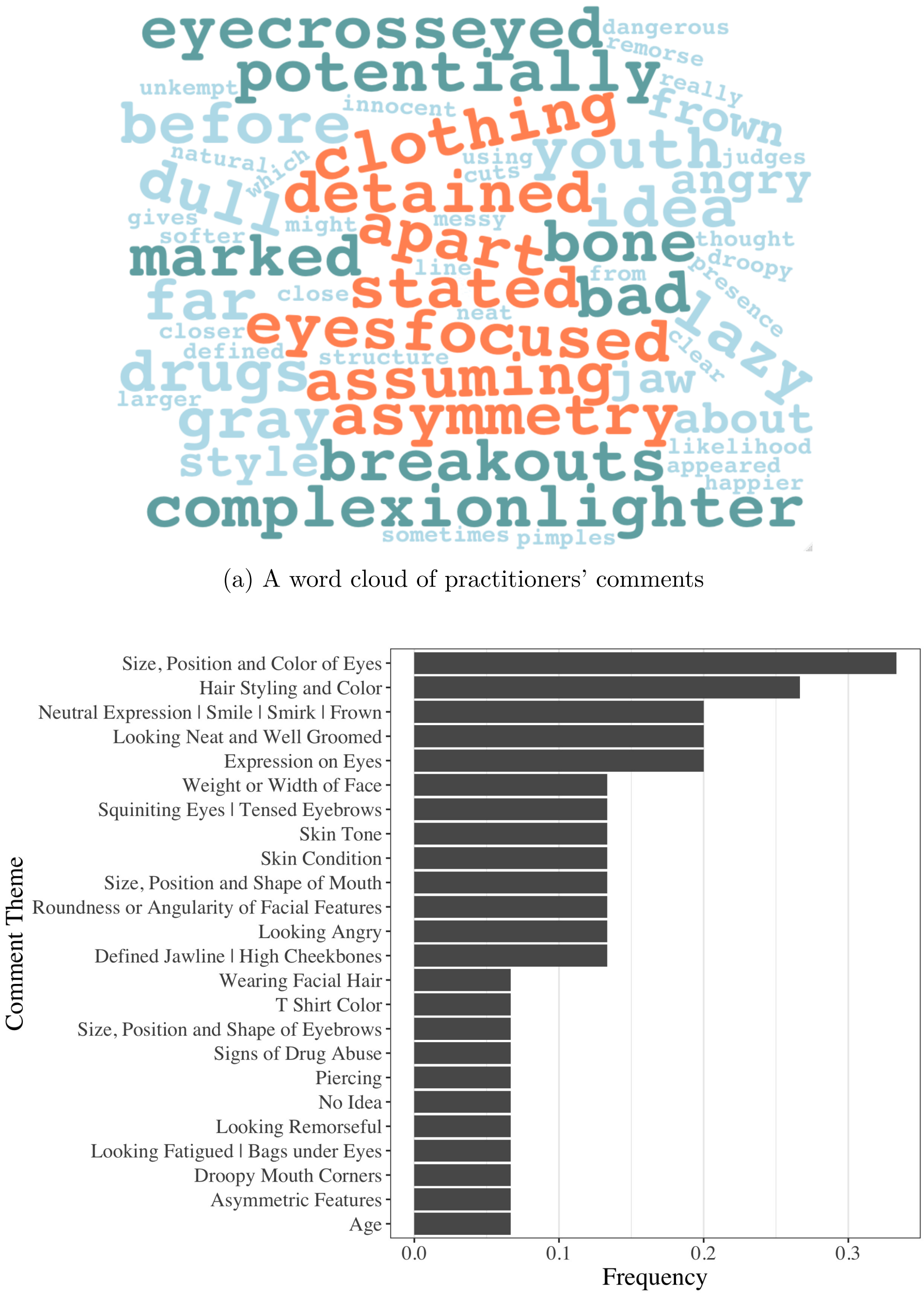


Figure A.XIX: Relative magnitude of the algorithm's discoveries on detention

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 children’s books,” *Quarterly Journal of Economics*, 2023, *138*, 2225–2285.

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,” *Harvard Law School Faculty Bibliography*, 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 Picard, "DISSECT: Disentangled Simultaneous Explanations via Concept Traversals," in "International Conference on Learning Representations" 2022.

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.

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.

Ludwig, Jens and Sendhil Mullainathan, "Machine Learning as a Tool for Hypothesis Generation," Technical Report, National Bureau of Economic Research 2023.

_ and _ , "Replication Data for: 'Machine Learning as a Tool for Hypothesis Generation'," 2023. Harvard Dataverse, https://doi.org/10.7910/DVN/ILO46V.

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.

Sheetal, Abhishek, Zhiyu Feng, and Krishna Savani, "Using machine learning to generate novel hypotheses: Increasing optimism about COVID-19 makes people less willing to justify unethical behaviors," *Psychological Science*, 2020, *31* (10), 1222–1235.

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*, 1175–1205.

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.