

WORKING PAPER · NO. 2023-28

Machine Learning as a Tool for Hypothesis Generation

Jens Ludwig and Sendhil Mullainathan

MARCH 2023

Machine Learning as a Tool for Hypothesis Generation*

Jens Ludwig
Sendhil Mullainathan[†]

February 15, 2023

Abstract

While hypothesis testing is a highly formalized activity, hypothesis generation remains largely informal. We propose a procedure that uses machine learning algorithms—and their capacity to notice patterns people might not—to generate novel hypotheses about human behavior. We illustrate the procedure with a concrete application: judge decisions. We begin with a striking fact: up to half of the predictable variation in who judges jail is explained solely by the pixels in the defendant’s mugshot—that is, the predictions from an algorithm built using just facial images. 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 factors implied by existing research (demographics, facial features emphasized by previous psychology studies), nor are they already known (even if just tacitly) to people or even experts. Though these results are specific, our procedure is general. It provides a way to produce novel, interpretable hypotheses from any high-dimensional dataset (e.g. cell phones, satellites, online behavior, news headlines, corporate filings, and high-frequency time series). A central tenet of our paper is that hypothesis generation is in and of itself a valuable activity, and hope this encourages future work in this largely “pre-scientific” stage of science.

Keywords: Hypothesis generation, algorithms, artificial intelligence

JEL codes: B4 (economic methodology), C1 (econometric and statistical methods and methodology: general)

*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. For valuable comments we thank Andrei Shliefer, 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, as well as seminar participants at Bocconi, Brown, Columbia, ETH Zurich, Harvard, MIT, Stanford, the University of California Berkeley, the University of Chicago, the University of Pennsylvania, the 2022 Behavioral Economics Annual Meetings and the 2022 NBER summer institute. For invaluable assistance with the data and analysis we thank Cecilia 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 of course our own.

[†]Ludwig: University of Chicago and NBER (jludwig@uchicago.edu); Mullainathan: University of Chicago and NBER (Sendhil.Mullainathan@chicagobooth.edu).

1 Introduction

Science is curiously asymmetric. New ideas are meticulously tested using data, statistics and formal models. Yet those same ideas originate in a notably less meticulous process involving intuition, inspiration and creativity. The asymmetry between how ideas are generated and how they are 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” (albeit 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 notice 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, at the same time, 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 inspiration is now machine readable: what was once solely mental data is increasingly becoming actual data.²

We suggest these changes can be leveraged to expand how we do science. Specifically, we suggest a procedure that applies machine learning algorithms on rich data sets to generate novel hypotheses. By “hypothesis” we mean a function $h(x)$ that relates explanatory variables x to some outcome, y . Our goal is to generate hypotheses that are novel, empirically plausible (supported by least a correlation of x and y), and falsifiable. We also want hypotheses that are human interpretable: the goal of science is to generalize knowledge to new contexts, which is only possible with hypotheses that make sense to people. Our proposed procedure is a cousin of the sort of exploratory data analysis (EDA): researchers looking at correlations to generate new hypothesis. The difference is the heavy lifting done by relies

¹The 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: discuss this literature and how our approach differs in the next section.

²See Einav and Levin (2014); Varian (2014); Athey (2017); Mullainathan and Spiess (2017); Gentzkow et al. (2019) and Adukia et al. (2021) on how these changes can affect economics.

on the idiosyncratic creativity deciding what correlations to try. In our approach, we use the machinery of supervised learning to systematically consider far more correlations than any human investigator could ever calculate and inspect—especially in high-dimensional data applications—and capitalize on its capacity

While our procedure is broadly applicable, we illustrate it in a concrete application: judicial decision-making: how pre-trial judges decide which defendants must go to jail versus go free awaiting trial. This is also a substantively interesting application in its own right because of the high stakes involved. But it is also intriguing because of the mounting evidence that judges make these decisions less than perfectly Kleinberg et al. (2018); Rambachan et al. (2021); Angelova et al. (2022). Though the decision by law is supposed to hinge on a prediction of the defendant’s risk (Dobbie and Yang, 2021), many other factors may end up mattering.³

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 has large explanatory power for who judges choose to detain: the defendant’s face. In fact, a predictor that uses *only* the pixels in the defendant’s mugshot explains from one-quarter to nearly one-half of the predictable variation in detention.⁴ Defendants whose mugshots 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 non-violent crimes is 4.8 pp. Notice what this finding is and is not. We are *not* claiming the mugshot predicts *defendant* behavior; that would be the long-discredited field of phrenology (Schlag, 1997). We are instead claiming the mugshot predicts *judge* behavior: how the defendant looks correlates strongly with whether the judge chooses to jail them or not.⁵

Has the algorithm found something new in the pixels of the mugshot 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 et al., 2018, 2020). When we control for age, gender, race, and skin color, and even the facial features suggested by previous psychology

³In practice, there are a number of additional nuances, as discussed in Subsection 3.1 and Appendix B.1.

⁴This is calculated for some of the most commonly-used measures of predictive accuracy, recognizing that different accuracy 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 et al., 2022). As a consequence, a predictive model of the judge can do better than the judge themselves (Kleinberg et al., 2018).

⁵In Section 4.2, we examine whether the mugshot’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 paper: uncovering hypotheses for later careful testing.

research (dominance, trustworthiness, attractiveness and competence), none of these factors (individually or jointly) meaningfully diminishes the algorithm’s predictive power.⁶

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 generating other controls, we take a different approach. We show mugshots to subjects asked to guess who the judge will detain, with incentives for accuracy. These guesses summarize the facial features people readily (even if implicitly) believe influence jailing. While subjects are modestly good at this task, the algorithm is much better. It remains highly predictive even after controlling for these guesses. The algorithm indeed 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 that allows intuitive communication. Our procedure produces pairs of synthetic mugshots that are “morphed” in the direction of higher or lower predicted detention likelihood. In principle we could have instead just shown subjects actual mugshots with higher versus lower predicted detention odds. But faces are so rich that, between any pair of actual mugshots, *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, we showed them to independent study subjects (M-turk or Prolific workers) in an experimental design to have them put a name on the algorithmic signal that they see. Specifically, we showed subjects pairs of morphed images and asked them 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 mugshots at about the same rate as they do when looking at “raw data,” that is, actual mugshots (modestly above the 50% random guessing mark), subjects quickly learn from these morphed images what the algorithm is seeing and reach an accuracy of nearly 70%. We then asked subjects to put words to the differences they see across images within each pair; that is, to name what they think are the key facial features the algorithm is relying on to predict judge decisions. Com-

⁶It is 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. The key point is that the algorithm has discovered a great deal more as well.

fortingly, 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 in fact used by the algorithm, a separate sample of subjects independently coded mugshots for this new feature. We then 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 the procedure produces meaningful communication and can be iterated.

What are the two discovered features? The first can be called “well-groomed” (e.g., tidy, clean, groomed, versus unkempt, disheveled, sloppy look), while 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: 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 pp (24% of the base rate) while the top versus bottom quartile difference in heavy-facedness is 7 pp (about 30% of the base rate). Both differences are larger than the 4.8 pp detention rate difference between those arrested for violent versus non-violent crimes.

Establishing whether these hypotheses are truly *causally* related to judge decisions is obviously beyond the scope of the present paper. But we nonetheless present a few additional findings that are at least suggestive. We show that these novel features do not appear to be simply proxies for other factors like substance use mental health, or socio-economic status. We also show that these hypotheses are novel even to people who work in the criminal justice system every day (practitioners from a public defender’s office and a legal aid society). And finally we carried out a lab experiment in which subjects are asked to make hypothetical pre-trial release decisions as if they were a judge. They are shown information about criminal records (current charge, prior arrests) along with mugshots that are randomly morphed. 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 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 paper about these specific findings. The paper is about developing 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, most of them related to meaningless intermediate concepts. But if we are to use algorithms for scientific discovery, our procedure cannot simply produce black box predictions. Non-inspectability is especially problematic when the data are rich and high-dimensional. This problem of interpretation is fundamental and still remains an active area of research.⁷ We resolve this problem by combining several different ingredients, some of which are borrowed, some of which are new. We rely on existing generative models to effectively synthesize faces; and use a new variant of gradient techniques to create our morphs. These component parts are combined in a way that is better suited for social science applications and distinct from other existing approaches (at least as far as we know).

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 are, like an ink blot, merely a canvas onto which we project our creativity.⁸ Put differently, a single research team lacks the kind of replicability we desire of a scientific procedure. To overcome this problem, we use independent (non-researcher) subjects. The fact that a sizable share of subjects all name the same discovery suggests 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 in two ways. Human subjects 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 such as ours. People also play a crucial role in curating the data the algorithm sees. Here, for example, we the researchers chose to include mugshots. So the creative acquisition of rich data is an important human input into this hypothesis generation procedure.⁹

⁷For reviews of the interpretability literature see Doshi-Velez and Kim (2017); Marcinkevičs and Vogt (2020); we discuss this literature and our approach below.

⁸Of course even if the hypotheses that are generated are the result of idiosyncratic creativity, this can still be useful. For example, Swanson (1986) and Swanson (1988) generated two novel medical hypotheses: the possibility that magnesium affects migraines; and that fish oil may alleviate Peychaud’s syndrome.

⁹But 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.

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 datasets 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, FOMC statements, Google searches, student essays, resumes, court transcripts, doctor notes, satellite images, housing photos, and medical images. Our procedure could, for example, raise hypotheses about what kinds of news lead to over- or under-reaction 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. So, beyond whatever the value might be of our specific procedure and empirical application, we hope these results also inspire greater attention to this traditionally “pre-scientific” stage of science.

2 A Simple Framework for Discovery

In this section we develop a simple framework to clarify the goals of hypothesis generation and how it differs from testing, how people currently generate hypotheses, how algorithms might help, our specific approach and how that differs from existing methods. A more formal treatment is in Appendix A.

2.1 The goals of hypothesis generation

Two obvious goals for hypothesis generation are *novelty* and *falsifiability* (Popper, 2005). The latter requirement that any new hypothesis is empirically testable rules out application of our procedure to many canonical computer science tasks, like image classification (does this image contain a cat?) Our procedure could in principle be used to suggest image features that the algorithm sees as cat-like. But whether those features are really what the algorithm sees as cat-like couldn’t be tested.¹⁰ To those two criteria we add two more: *interpretability* and *empirical plausibility*.

We care about interpretability because science is in large part 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,

¹⁰Thanks to an anonymous reviewer for this point.

nothing could be said about a new set of defendants. In contrast an interpretable hypothesis like “skin color affects detention” has implications not only for other samples of defendants but even 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). Interpretable descriptions let us generalize to novel situations, in addition to being easier to communicate to key stakeholders and lending themselves to interpretable solutions.

By empirically plausible we mean there exists some correlation between $h(x)$ and y . Of course our ultimate aim is to uncover causal relationships. But causality can only be known *after* causal testing. What might we want to see in deciding an idea is worth testing? Many true hypotheses may not be visible in raw correlations. Those can only be identified with background knowledge (e.g., theory). Other procedures beyond ours would be required to surface those. Our focus is on searching for true hypotheses that *are* visible in raw correlations. Of course not every correlation will be a true hypothesis, but generating such hypotheses and then invalidating them is in and of itself a valuable activity. Debunking spurious correlations has long been one of the most useful roles of empirical work. And understanding what confounders produce those correlations can also be useful.

2.2 Human hypothesis generation

A key feature of much human hypothesizing from data seems to be our tendency to notice *contrasts*.¹¹ Humans essentially look at a subset of all data and look for something that differentiates positive and negative cases with respect to some outcome. We typically focus on differences that hold *in these data*, rather than “out of sample.”

Human hypothesis generation has the advantage of generating hypotheses that are interpretable: by construction the ideas that human beings come up with are understandable by human beings. But as a procedure for generating new ideas, human creativity has the drawback of often being idiosyncratic and so 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 even the same person may notice different things at different times (Kahneman et al., 2022). Nor are we able to introspect and understand why we notice specific things those times we do notice them.¹²

¹¹Of course the psychology (and sociology) of this process is enormously complex; see for example Langley et al. (1987). We are just trying to capture just some high-level features.

¹²This is related to what Autor (2014) called “Polyani’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 of people to access their own cognition see for example Wilson (2004) and Pronin (2009).

Nor will human-generated hypotheses necessarily be empirically plausible. The intuition is related to “over fitting”: even with no noise in y , there is randomness in which observations are in the data we inspect. 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.

Consider a simple example: Suppose $x = (x_1, \dots, x_k)$ 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 non-plausible 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 still remains.

2.3 Algorithmic Hypothesis Generation

How might algorithms help with this activity? Supervised learning tools in machine learning are designed to generate predictions, $m(x)$, that are accurate in *new* (out-of-sample) data.¹³ That is, algorithms generate hypotheses that are empirically plausible by construction.¹⁴ 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 EKG (Mullainathan and Obermeyer, 2022), or that a piece of machinery is about to break (Mobley, 2002).

The challenge is that most machine learning prediction functions 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 dataset) and we used a predictor such as OLS (or LASSO), then we could just read the hypotheses from the non-zero coefficients: which variables are significant? Even in that case, interpretation is challenging because machine learning tools, built to generate accurate predictions rather than to apportion explanatory power across explanatory variables, yield coefficients that can be unstable across realizations of the data (Mullainathan and Spiess,

¹³Some canonical references include Hastie et al. (2009), Breiman (2001), Jordan and Mitchell (2015) and Breiman et al. (2017). For discussions about how machine learning connects to economics see Belloni et al. (2014), Varian (2014), Mullainathan and Spiess (2017), Athey (2018) and Athey and Imbens (2019).

¹⁴Of course there is not always 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.

2017).¹⁵ 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. Moreover the models are defined on granular inputs that have no particular meaning: if we knew the algorithm weighted a particular pixel, what have we actually learned? In these cases, the estimated model m is itself not interpretable. 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 that 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.

Our approach is to construct counterfactual pairs of synthetic mugshot images that differ in their predicted detention odds, but are as similar as possible along other dimensions that are irrelevant to judge detention decisions. This requires not just an algorithmic model of y but an algorithmic model of the data distribution of faces within pixel space, p . That model of p frees the algorithm from being constrained by the particular set of pairs found in the data set of actual mugshots, and allows for the construction of entirely new data points that differ only along relevant dimensions, since we can use the prediction $m(x)$ to choose the new matching points. That is, for any given data point this procedure allows us to construct new data points that answer the counterfactual question: How would this point be different if it had a higher or lower $m(x)$ value?

With these synthetic “morphed” counterfactuals in hand, we can then harness two key human capacities: the ability of people to notice differences in otherwise similar observations; and the unique human knowledge of what is a meaningful hypothesis. Because people are not looking at actual data, they are effectively *naming* $m(x)$. They are projecting the algorithm into their own language—the set of hypotheses that are comprehensible. As a result, mechanically, this produces comprehensible hypotheses.

At the same time, because $m(x)$ is known to have signal for y , relative to having people look at pairs of actual instances this morphing procedure is more likely to produce empirically plausible hypotheses. Left on their own, people seek to identify *any* differences across data points that differ in y . Because they only have one particular data set, that approach can lead to people to see differences in raw data that hold in that sample but not others. But with our morphing procedure, humans are now looking at morphed instances to spot differences in

¹⁵The intuition here is quite straightforward: If two predictor variables are highly correlated, the weight that the algorithm puts on one versus the other can change from one draw of the data to the next depending on the idiosyncratic noise in the training dataset, but since the variables are highly correlated the predicted outcome values themselves (hence predictive accuracy) can be quite stable.

$m(x)$, and we know $m(x)$ is a reliable out-of-sample predictor.¹⁶ Relatedly, the matching of nearest neighbors in our morphing procedure reduces the curse of dimensionality. Morphed synthetic pairs will now have fewer plausible candidates for what may be different between them exactly because we have ensured they differ as little as possible.¹⁷

Finally, it is worth noting a few additional advantages of algorithmic hypothesis generation relative to human generation. First, though it is not explicit here, given a set of *known* hypotheses, we can orthogonalize with respect to those dimensions to ensure that the algorithm is producing something novel.¹⁸ Second, other methods of producing hypotheses (observation, conversation, introspection) may produce theories that are hard to measure in data. By construction, our procedure only produces hypotheses that are measurable.

2.4 Related Methods

Concurrent with our own work, a few papers in computer science have developed related methods mostly for application to what we call “automation” tasks: to automate with AI something that a human can do nearly perfectly like image classification (Is this image of a dog or a cat?) Interpretability techniques are needed for those applications in large part to ensure the algorithm is not picking up on spurious signal. If for example every dog in a dataset was photographed outdoors and every cat indoors the algorithm might key on background in making its predictions, which could then lead to error in more representative data. But those methods are not so well suited for applications like predicting human behavior.

For example, like us, Lang et al. (2021) also combines a predictive supervised learning model $m(x)$ with a generative model for images $p(x)$, specifically a generative adversarial network, or GAN (Goodfellow et al., 2014b). But their approach assumes the latent space of faces for the GAN is composed of disentangled attributes that cleanly separate $y = 1$ from $y = 0$ cases. This approach can work with automation tasks where the algorithm’s predictive accuracy can be quite high, but their key assumption is less likely to hold when trying to use an algorithm to learn predictors of human behavior (like judge decisions), for which predictive accuracy is almost always far lower.¹⁹ To make this concrete, for automation tasks like image

¹⁶By way of intuition, this is why we are usually better off interpreting a regression than individual data points that go into the regression.

¹⁷An analogy with OLS provides an easy intuition for why this helps: a regression is in effect a way to calculate in-sample correlations, and the standard errors tell us whether those correlations are real or due to sampling noise. Adding controls lowers standard errors because lowering the residual noise makes it more likely that in-sample correlations are more likely to hold out-of-sample. That is the same effect here: by matching on $m(x)$ instead of y , we are reducing residual noise and increasing the chance that noticed patterns are genuine ones.

¹⁸In our particular application, we do not do this, in part because we are curious to explicitly examine algorithms’ capacity to rediscover known hypotheses.

¹⁹Relatedly, Narayanaswamy et al. (2020) develop a procedure that is able to ensure new synthetic instances stay on the data distribution within the context of outcomes that can be predicted very well; in applications like ours, where the outcome is human behavior and so intrinsically difficult to predict, their approach would

classification the predictive accuracy (AUC) can be on the order of 0.99; that is, in a dataset split evenly between $y = 1$ and $y = 0$ cases there is a 99% chance that a randomly selected positive case will have a higher predicted value than a random negative case. In contrast, our model that predicts judge decisions using the face only achieves an AUC of .625. These high levels of unpredictability - common to nearly all social science applications - requires different approaches.

In addition many of these existing papers in the interpretability literature use methods that only work with a very particular sort of generative model, such as a StyleGan (Karras et al., 2019) or a variational auto-encoder (Miller et al., 2019). In contrast, our approach is agnostic to the particular generative model used, including those that use other types of high-dimensional data, not just images, and so is more generally applicable.

Our approach is also part of a growing literature that aims to integrate machine learning into the way science is conducted. A very 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 has been very hard to predict the final shape of a protein. In such cases, machine learning has provided fundamental breakthroughs, in effect by making very hard-to-compute outcomes computable in a feasible timeframe.²⁰

Progress has been far more limited with applications where the relationship between x and y is *unknown* (what we call “open world” problems), like human behavior. First, machine learning here has been useful at generating unexpected findings, but that are not hypotheses themselves. For example, 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.²¹ Second, machine learning has also been used to explore *investigator*-generated hypotheses, such as (Mullainathan and Obermeyer, 2022) who examine “physicians suffer

have difficulty ensuring new morphed face images actually still look like faces. Ghandeharioun et al. (2021) develop a procedure that is unlikely to work when the predictive model draws on attributes that interact in their influence on the classification outcome (for example, if the effect of long hair on the odds of some outcome is different for men versus women). Liu et al. (2019) develop a procedure that seeks to generate new synthetic instances that simultaneously maximize the difference in the class prediction, $m(x)$, while minimizing the difference in pixel values between the original and new morphed synthetic image. But image pixel values are not the same as visual image similarity to humans. So their procedure can lead to morphed counterfactuals that are close in pixel space but do not look visually similar to study subjects (that is, do not hold visually constant other key features of the image that are irrelevant to the prediction). Moreover their procedure has been shown to work only on small (low resolution) images: they test their procedure on the MNIST character dataset where the image resolutions are 28×28 , and on the CelebA dataset where resolution is 178×218 , while our mugshot images are substantially larger (400×480).

²⁰Examples of applications of this type include Carleo et al. (2019), He et al. (2019), Davies et al. (2021) and Jumper et al. (2021), and Pion-Tonachini et al. (2021).

²¹As other examples, researchers have found that retinal images alone can unexpectedly predict gender of patient or macular edema (Korot et al., 2021; Narayanaswamy et al., 2020)

from limited attention when diagnosing patients.” 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 then inspected those algorithms to produce their insights. Similarly, (Kleinberg et al., 2018; Sunstein, 2021) build algorithmic models of judges and examine that model to generate hypotheses.²² Our proposal builds on these papers. Rather than focusing on generating an insight for a specific application, it suggests a procedure that can be broadly used for many applications. And, importantly, our procedure does not rely on *researcher* inspection of algorithmic output. As noted above, when an expert researcher with a track record of generating scientific ideas uses some procedure to generate an idea, how do we know whether if 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.

3 Application and Data

3.1 Judicial Decision Making

While our procedure is broadly applicable, we illustrate it through a specific application to the US 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 US) that are recorded in data, and, increasingly, high-dimensional data that can also be used to model those choices, such as mugshot images, police body-worn cameras, and text from arrest reports or court transcripts—are shared with a variety of other applications.

Our specific focus is on pre-trial 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 9-12 months. Jail affects people’s families, their livelihoods, and the chances of a guilty plea (Dobbie et al., 2018). On the other hand, someone who is released could potentially re-offend.²³

While pre-trial decisions are by law *supposed* to hinge on the defendant’s risk of flight or re-arrest 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 et al., 2021),

²²Closest to the present application is Miller et al. (2019), which morphs EKG output but stops at the point of generating realistic morphs and does not carry this all the way through to generating interpretable new hypotheses.

²³Additional details about how the system works are in Appendix B.

partly because judges over-weight the charge for which people are arrested (Sunstein, 2021). Judge decisions can also depend on extra-legal factors like race (Arnold et al., 2018, 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 et al., 2016), and if the hearing is on the defendant’s birthday (Chen and Philippe, 2020).

These studies each test a hypothesis 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.

3.2 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 US 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 US as a whole (56.6% versus 75.8%).²⁴ We rely on three sources of administrative data:²⁵

- The *Mecklenburg County Sheriff’s Office (MCSO)* publicly posts arrest data for the past 3 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 pre-trial 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 “mugshots”),²⁶ 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 described below. In Figure I, we give readers a sense of what these mugshots look like but with two important caveats. First, given concerns about how the over-representation 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

²⁴For Black non-Hispanic the figures for Mecklenburg County versus the US were 33.3% versus 13.6%. See <https://www.census.gov/programs-surveys/sis/resources/data-tools/quickfacts.html>

²⁵Details on how we operationalize these variables are in Appendix B.

²⁶The mugshot 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>)

for non-Hispanic white males. Second, out of sensitivity to actual arrestees, we do not wish to display *actual* mugshots (which are available at the MCSO’s website²⁷). Instead, the paper only shows mugshots that are *synthetic*, generated using GANs as described in Section 5.2.

These data capture much of the information the judge has available at the time of the pre-trial 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 mugshot (which the algorithm uses) is taken not long before the pre-trial 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 arrest report from the police, and what happens in court. While pre-trial 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 themselves usually have their lawyers speak for them and so 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 (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 $N = 23,138$ cases, constructed by taking arrests through July 17, 2019, grouping arrests by arrestee,²⁸ 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 OOS performance in this draft, consisting of the remaining 30% in the combined training-plus-validation data frame.
- A *hold-out set* of $N = 19,009$ cases we use to report OOS performance after the paper is accepted (to avoid inadvertently overfitting the OOS data as we respond to seminar or referee suggestions, etc.). This consists of the $N = 13,728$ 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.

Descriptive statistics are shown in Table I. Relative to the county as a whole, the arrested population substantially over-represents men (78.7%) and Black residents (69.4%). The average age of arrestees is 31.8 years. Judges detain 23.4% of cases, and in 25.1% of arrests the person is re-arrested before their case is resolved (so 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 0.10 (see Appendix Table A.II). A multivariate analysis of variance test of the joint null hypothesis

²⁷<https://mecksheriffweb.mecklenburgcountync.gov/>

²⁸We 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.

that the training-validation differences for all variables are all zero yields $p = 0.42$.²⁹

3.3 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 mugshots. More details are in 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.³⁰ Because demographics tend to be easy 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 now collected 10 labels for each image; we see additional labels add little information.³¹ Another data quality check comes from the fact that the distributions of skin color ratings do systematically differ by defendant race (Appendix Figure A.IV).

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: (1) *Trustworthiness* and (2) *Dominance*. We also collected data on two other facial features shown to be associated with real world decisions like hiring or who to vote for: (3) *Attractiveness* and (4) *Competence* (Frieze et al., 1991; Little et al., 2011; Todorov and Oh, 2021).

We asked subjects to rate images for each of these psychological features on a 9-point scale. Because these psychological features may be less obvious than demographic features,

²⁹To account for the non-independence of observations, i.e., arrests, across individual arrestees, we re-shape the data into “long” format—a row for each arrest-and-variable—then adjust the degrees of freedom in the usual ANOVA F-test for the number of separate arrestees in the data set.

³⁰For an example HIT task see Appendix Figure A.III.

³¹For 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).

we collected three labels per training-data set image and five per validation-data set image.³² There is substantial variation in the ratings that subjects assign to different images for each of these features (see Appendix Figure A.VI). And the ratings from different subjects for the same feature and image are highly correlated: inter-rater reliability measures (Cronbach’s α) range from 0.92 to 0.97 (Appendix Figure A.V), similar to those reported in studies like Oosterhof and Todorov (2008).³³ The information gain from collecting more than a few labels per image is modest.³⁴ For summary statistics see 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 mugshot out of a given pair would be detained, with images in each pair matched on gender, race and five-year age brackets.³⁵ We incentivized study subjects for correct predictions and give them feedback over the course of the 50 image pairs that they see in order to facilitate learning. We treat the first 10 responses per subject as a “learning set” that we exclude from our analysis.

4 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 here 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.

4.1 What drives judge decisions?

We begin by predicting judge pre-trial 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 ; and for the unstructured data (raw pixel values from the mugshots) we train a convolutional neural network (CNN), m_u . Each model returns an estimate of y (a predicted detention probability) for a given x . Because these initial steps of our procedure use standard machine learning methods, we relegate their discussion to the appendix.

We pool the signal from both models to form a single weighted-average model $m_p =$

³²We 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.

³³For 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.

³⁴See Appendix Figure A.VII, 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.

³⁵For an example see Appendix Figure A.VIII.

$[\hat{\beta}_s m_s + \hat{\beta}_u m_u]$ using a so-called “stacking” procedure where the data are used to estimate the relevant weights.³⁶ Combining structured and unstructured data is an active area of deep-learning research, often called fusion modeling (Yuhas et al., 1989; Lahat et al., 2015; Ramachandram and Taylor, 2017; Baltrušaitis et al., 2019). We have tried several of the latest fusion architectures; none improve on our ensemble approach.

Figure II shows that judge decisions do indeed have some predictable structure. We report predictive performance as the area under the receiver operating characteristic (ROC) 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.

What is the algorithm using to make its predictions? Figure II shows that just a single type of input captures a sizable share of the total signal: the defendant’s face. The algorithm built using *only* the mugshot image, $m_u(x)$, has an AUC of 0.625. Since an AUC of 0.5 represents random prediction, in AUC terms the mugshot accounts for $(0.625 - 0.5)/(0.780 - 0.5) = 44.6\%$ 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 so 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, but R^2 , unlike AUC, *also* captures how close predictions are to observed outcomes (*calibration*).³⁷ 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 ($0.03/0.11 = 27.3\%$) 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 Figure III, the difference in detention rates equals 4.8 pp for those arrested for violent versus non-violent

³⁶We use the validation data set to both estimate $\hat{\beta}$ and then also evaluate the accuracy of m_p . While 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.

³⁷The MSE for a linear probability model’s predictions is related to the “Brier score” (Brier et al., 1950). For a discussion of how this relates to AUC and calibration, see Murphy (1973).

crimes, 10.2 pp for males versus females, and 4.3 pp 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 mugshot-algorithm’s predictions, $m_u(x)$, differ by 20.4 pp.

In what follows, we seek to understand more about the mugshot-based prediction of the judge’s decision, which we refer to simply as $m(x)$ in the remainder of the paper.

4.2 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 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 re-arrest against the algorithm’s mugshot-based detention prediction. That yields a statistically significant relationship: The coefficient (and standard error) for the mugshot equals 0.6127(0.0461) with no other explanatory variables in the regression versus 0.5151(0.0527) with all the explanatory variables (as in the final column, Table 3). 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 re-arrest rates may correlate with facial characteristics simply due to measurement bias. The problem created by having measures of re-arrest only for released defendants is that if judges have access to private information (defendant characteristics not captured by our dataset), 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 re-arrest risk (Kleinberg et al., 2018).

With these caveats in mind, at the very least we can perform a bounding exercise. We created a predictor of re-arrest risk (see Appendix C) and then regress judges’ decisions on predicted re-arrest risk. We find that a 1-unit change in predicted re-arrest risk changes judge detention rates by 0.6103 (standard error 0.0213). By comparison, we found that a 1-unit

change in the mugshot (by which we mean the algorithm’s mugshot-based prediction of the judge detention decision) changes judge detention rates by 0.6963 (standard error 0.0383; see column 1 of Table III). That means if the judges were reacting to the defendant’s face *only* because the face is a proxy for re-arrest risk, the difference in re-arrest risk for those with a 1-unit difference in the mugshot would need to be $0.6963/0.6103 = 1.141$. But when we *directly* regress re-arrest against the algorithm’s mugshot-based detention prediction we get a coefficient of 0.6172 (standard error 0.0460). Clearly $0.6172 < 1.141$; that is, the mugshot does not seem to be strongly related enough to re-arrest risk to explain the judge’s use of the mugshot in making detention decisions.³⁸

Of course this leaves us with the second problem with our data mentioned above: we only have crime data on the released. It is possible the relationship between the mugshot and risk could be very different among the 23.3% of defendants who are detained (which we cannot observe). Put differently, while the mugshot-risk relationship among the 76.7% of the defendants who are released is 0.6172, what we really want to know is the mugshot-risk relationship among *all* defendants, which equals $(0.767 \cdot 0.6172) + (0.233 \cdot X)$. For this mugshot-risk relationship among all defendants to equal 1.141, X would need to be 2.985; that is, the relationship between the mugshot and re-arrest risk would need to be nearly five times as great among the detained defendants as among the released. This would imply an implausibly large effect of the mugshot on re-arrest risk relative to the size of the effects on re-arrest risk of other defendant characteristics.³⁹

In addition, the results from Section 6.2 call into question that these characteristics are well-understood proxies for risk. As we show there, experts who understand pre-trial (public defenders and legal aid society staff) do not recognize the signal about judge decision-making that the algorithm has discovered in the mugshot. 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 rather reflects errors in the judicial decision-making process. Of course, that presumption is not essential for the rest of the paper, which asks: what exactly has the algorithm discovered in the face?

³⁸Note how this comparison helps mitigate the problem that police arrest decisions could depend on a person’s face. When we regress re-arrest against the mugshot, 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 re-arrest risk, some of the variation across defendants in re-arrest 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.

³⁹The average mugshot-predicted detention risk for the bottom and top quartiles equal 0.127 and 0.332; that difference times 2.985 implies a re-arrest risk difference of 59.3 pp. By way of comparison, the difference in re-arrest risk between those who are arrested for a felony crime rather than a less-serious misdemeanor crime is equal to just 7.8 pp.

4.3 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 see stereotyping on the basis of personal attributes like gender (Avitzour et al., 2020), age (Dahl and Knepper, 2020; Neumark et al., 2016) and race or ethnicity (Bertrand and Mullainathan, 2004; Arnold et al., 2018, 2020; Goncalves and Mello, 2021; Hoekstra and Sloan, 2020; Fryer Jr, 2020). Is the algorithm just rediscovering known determinants of people’s decisions, or is the algorithm discovering something new? We try to answer this question in two ways. We first ask how much of the algorithm’s *predictions* can be explained by various 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 these 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.⁴⁰

Columns (1) to (3) of Table II show the relationship of the algorithm’s predictions to demographics. The algorithm’s predictions vary enormously by gender (men have predicted detention likelihoods 11.9 pp higher than women), somewhat less so by age,⁴¹ and also 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 pp 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 pp lower predicted likelihood of detention compared to whites.⁴²

Column (4) of Table II 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 turn out to help explain the algorithm’s predictions of judges’ detention decisions: People judged by independent raters to be 1 standard deviation more attractive, competent or trustworthy have lower predicted likelihood of detention equal to 0.51, 0.83 and 0.41 pp, respectively, or 2.2%, 3.6% and 1.8% of the base rate.⁴³ Those whom subjects

⁴⁰In 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 Appendix Tables A.VI, A.VII, and A.VIII.

⁴¹With 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 ($10 \times 0.0006 = 0.6$ pp higher likelihood of detention, versus 11.9 pp).

⁴²Appendix Table A.V shows that Hispanic ethnicity, which we measure from subject ratings from looking at mugshots, is not statistically significantly related to the algorithm’s predictions. Column (2) of Table II showed that conditional on gender, Black defendants have a slightly higher predicted detention odds than white defendants (0.3 pp), but this is not quite significant ($t = 1.3$). Column (1) of Appendix Table A.V 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 pp) as well as the difference’s statistical significance ($t = 2.2$).

⁴³This comes from multiplying the effect of each 1 unit change in our 9-point scale associated, equal to

judge are 1 standard deviation more dominant-looking have a higher predicted likelihood of detention of 0.35 pp (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. We can then 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 mugshots matched by gender, race and 5-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 chances subjects take this exercise seriously.⁴⁴ 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 socio-economic 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). Subjects guess which mugshot is more likely to be detained at a rate of 51.4%, which is statistically significant from the random-guessing 50% threshold. When we regress the algorithm's predicted detention rate against these subject guesses, the coefficient is 3.99 pp, equal to 17.1% of the base rate.

Together, the findings in Table II are somewhat remarkable. The only input the algorithm had access to was the raw pixel values of each mugshot, yet it has re-discovered findings from decades of previous social science research and human intuition. At the same time, 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, of course, does not tell us much. It is possible that the remaining variation is prediction error—components of the prediction that do not explain actual judges' decisions.

0.55, 0.91 and 0.48 percentage points, respectively, with the standard deviation of the average label for each of these psychological features for each image, which equal 0.923, 0.911 and 0.844, respectively.

⁴⁴As discussed in 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.

In Table III, we therefore test whether the algorithm uncovers any additional signal above and beyond 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), barely changes 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.⁴⁵

5 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 will have simply replaced one black box hypothesis-generation procedure (human creativity) with another (the algorithm). In what follows we first demonstrate how existing methods like saliency maps cannot solve this challenge in our application, and then discuss our solution to that problem.

5.1 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 $m(x)$ 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 $\nabla m(x) = \frac{dm}{dx}(x)$. This lets us determine, at any given input value, what change in the input vector maximally changes the prediction.⁴⁶ The idea of gradients is useful for image classification tasks because it allows us to tell which pixel image values are most important for changing the predicted outcome.

⁴⁵Table 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 (col. 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.” So controlling for these factors directly adds additional signal.

⁴⁶Imagine a linear prediction function like $m(x_1, x_2) = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. If our best estimates suggested $\hat{\beta}_2 = 0$, the maximum change to the prediction comes from incrementally changing x_1 .

For example, a widely used method known as “saliency maps” uses gradient information to highlight which specific pixels are most important for predicting the outcome of interest (Baehrens et al., 2010; Simonyan et al., 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 wind up highlighting pixels around, say, the cat’s head: that is, what is most cat-like is not the tail, paws or torso, but the eyes, ears and whiskers. Unfortunately with many economically-relevant behaviors, they depend on complicated functions of the entire image. For example in our empirical application, as we demonstrate below, the saliency map highlights almost the whole face.

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, say, 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 *under-communicates* with humans.

To test the ability of saliency maps to help with our application, we focused for starters 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 mugshot.⁴⁷ For a representative image, as in the top left of Figure IV, we can then highlight which pixels are most important for predicting age, shown in the top right.⁴⁸ 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.⁴⁹

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

⁴⁷As discussed in Appendix C, when we trained our CNN to predict judges’ detention risk using the image, we actually trained a multiple outcome or “multi-head” CNN that simultaneously also predicts gender, race and age (defined as above versus below our sample’s median age). This is the basis for our $m(x)$ predictor for age in this discussion.

⁴⁸As noted above, to avoid contributing to the stereotyping of minorities in discussions of crime, in our exhibits we show images for non-Hispanic white males, although in our HITs we use images representative of the larger defendant population.

⁴⁹Because they are often used in statistics books to teach principal components analysis, “eigenfaces” are another candidate solution that might come to mind (Sirovich and Kirby, 1987). These aim to provide a low-dimensional representation of face space. But they do not provide much useful insight for our application, as seen from the eigenfaces calculated using our dataset (described below) shown in appendix figure A.I. In some sense it is obvious why this would not work for our application: 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.

different predicted outcome, what we call “morphing.” This new image 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 V summarizes our semi-automated hypothesis-generation pipeline. (For more details see Appendix C.) The benefit of morphed not actual mugshot images is to isolate the differences across faces that matter for the outcome of interest. Morphing, by reducing noise, also reduces the risk of spurious discoveries.

Figure VI illustrates how this morphing procedure works in practice, and also 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 mugshot-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 mugshot image. We could then 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. So simply following the gradient will usually take us off the data distribution of face images, as illustrated abstractly in the top panel of Figure VI. What this means in practice is shown in the bottom-left panel of Figure IV: The result is an image that has a different predicted outcome (in the figure, 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.

5.2 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.⁵⁰ Specifically, we use generative adversarial networks (GANs), originally introduced to generate realistic new images for a variety of tasks (see for example Goodfellow et al. (2014b)).⁵¹

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.

⁵⁰Modeling $p(x)$ through a supervised learning task would involve assembling a large set of images, having subjects label each of these images for whether they contain a realistic face, then predict 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.

⁵¹See Kaji et al. (2020), Athey et al. (2021) and Athey et al. (2022) for some recent uses of GANs in economics problems.

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. 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 data distribution of faces. (Additional discussion of GANS in general, and how we construct our GAN specifically, are in Appendix C.)

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 Appendix C.3 for details). A more qualitative way to judge the quality of our GAN comes from visual inspection; some examples of synthetic face images are in Figure I. Most importantly, the GAN we build (and GANs in general) are not generic. They are specific. They do not generate “faces,” but instead seek to match the distribution of pixel combinations in the training data. So, for example, our GAN trained using mugshots would never generate generic Facebook profile photos, or a celebrity headshot.

Figure VI 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 (mugshots), 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 to not just 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? There is in principle 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 varies 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 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 mugshots). In practice, though, there is a limit: what can change depends on how rich and expressive the estimated GAN itself is. If the GAN, for example, 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.

Of course there is no guarantee that in any given application the classifier $m(x)$ need find novel signal for the outcome y . Nor need it be 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, we leave to future work. Whether our procedure works for the specific application of judge decisions is the question to which we turn next.⁵²

5.3 Validating the Morphing Procedure

We return now 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. Now 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 IV. 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.

To figure out if algorithm-human communication occurs, we run these age-morphed image pairs through our experimental pipeline (Figure V). 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 carrying out this study), but rather a sample of external, independent study subjects. In our application, we use Prolific workers (see Table A.III). Reliability or replicability would be indicated by the level of agreement in the subject responses: lots of subjects see and articulate the same thing.

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 within each pair differ on some hidden dimension but did not tell them what that was.⁵³ We asked subjects to guess which image

⁵²Some 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 it’s classification for the outcome y , so that mugshots 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 for example Szegedy et al. (2013) and Goodfellow et al. (2014a). In practice such instances are quite rare in nature, indeed so rare they usually occur only if intentionally (maliciously) generated. This seems unlikely (at least at present) for many of the questions economists study, but that could of course change over time.

⁵³Appendix Figure A.XI 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

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 within 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 lower-case characters, and removing stop words. We then gave 3 research assistants not otherwise involved in the project these responses and asked them to create their own categories that together would capture all the responses (see Appendix A.XII). 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 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.⁵⁴

5.4 Understanding the Judge Detention Predictor

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

The top panel of Figure VII 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 VIII.

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,⁵⁵ and gave them feedback on the right answer. Figure IX 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

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.

⁵⁴In principle this semantic grouping could be carried out in other ways, for example, with automated procedures involving natural language processing.

⁵⁵See Table A.III for a high-level description of this human intelligence task, and Appendix Figure A.XIII for a sample of the task and the subject instructions).

when subjects look at pairs of actual mugshots (where accuracy is 51.4% across the final 40 mugshot-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 pp after the 20th morphed image pair and reached 67% after 30 image pairs. Compared to looking at actual mugshots, the morphing procedure accomplished its goal of making it easier for subjects to see what in the face matters most for detention risk.

We then 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 X, we present a histogram of individual tokens (cleaned words from worker comments) in “word cloud” form, where word size is approximately proportional to frequency.⁵⁶ Some of the most common words are “shaved,” “cleaner,” “length,” “shorter,” “moustache” and “scruff.” To form semantic categories, we use a procedure similar to what we describe above for our validation exercise for the known feature of age.⁵⁷ Grouping tokens into semantic categories, 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 X).

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 (M-Turk workers) to label the 32,881 mugshots in our combined training and validation data sets for how well-groomed each image was perceived to be on a 9-point scale (see Table A.III).⁵⁸ For datasets of our size these labeling costs 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

⁵⁶We drop every token of just 1 or 2 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.”

⁵⁷We enlisted 3 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 RA mapped each subject comment to 5% of semantic categories on average, if the RA mappings were totally uncorrelated, we would expect to see agreement of at least two RA categorizations about 5% of the time. What we actually see is if one research assistant made an association, 60% of the time another RA would make the same association. We assign a comment to a semantic category when at least two of the RAs agree on the categorization.

⁵⁸This 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. Sampling labels from different workers on the same image, these ratings have a correlation of 0.14.

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 pp, 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 .

5.5 Iteration

Our procedure is also 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 not just on how important each feature is in explaining the judge’s detention decision, but also 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 well-groomed.⁵⁹ But iterating the procedure raises several technical challenges. To see these technical challenges, consider first 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 towards higher predicted detention risk $m(x)$, but orthogonal to $h_1(x)$,
- 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.⁶⁰ That means we

⁵⁹It turns out that skin tone is another feature that winds up being 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 of the features we orthogonalize against, well-groomed plus skin tone.

⁶⁰To 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

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).⁶¹ The errors in the process accumulate as we take many morphing steps. Worse, 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.

So we instead 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 ($y_i = 1$), we find a released observation ($y_j = 0$) with $h_1(x_i) = h_1(x_j)$. Because within that training data set y is now orthogonal to $h_1(x)$, we can use the gradient of the orthogonalized judge predictor to move in GAN latent space to create new morphed images that have different detention odds but are similar with respect to well-groomed.⁶² We call these “orthogonalized morphs,” which we then feed into the same experimental pipeline shown in Figure V.⁶³ 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.

The results of this orthogonalized image-morphing procedure are illustrated in Figure XI, with additional examples in Figure XII. The top panel shows an image pair, with intermediate steps and their predicted detention risk shown underneath. 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, ranging from 0.07 up to 0.27. We can also see that the changes in facial features across morphed images are notably different from those we see in the first iteration of morphs in Figure VII. From these examples, it appears possible that orthogonalization may be slightly imperfect; sometimes they show subtle differences in “well groomed” and perhaps age.

Still, there is a salient new signal: when presented to subjects they name a second facial feature, as shown in Figure XIII. 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 XIII shows that some of the most common terms reported

final morph requires hundreds of such intermediate morphing steps, the whole process would be so time- and resource-consuming as to be infeasible.

⁶¹While 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 out-of-sample 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 mugshots, are themselves noisy, which introduces irreducible error.

⁶²For additional details see Appendix Figure A.IX and Appendix C.

⁶³There are a few additional technical steps required, discussed in Appendix C. 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 we use to generate independent labels for that new hypothesis or facial feature, see Table A.III.

by subjects include “big,” “wider,” “presence,” “rounded,” “body,” “jaw” and “head.” When we ask independent RAs 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 M-turk workers, see 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 (column 1 of Table V). 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 pp, 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 taken together still only explain 27% of the variation in the algorithm’s predictions of the judges’ decisions. So long as there is residual variation, the hypothesis-generation crank could be turned again and again. But since our goal is not to fully explain judges’ decisions but rather to illustrate that the procedure works and is iterable, we leave this for future work (ideally done on data from other jurisdictions as well).

6 Evaluating These New Hypotheses

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

6.1 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 is, after all, 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 $m(x) - y$.

But 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 mugshot-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 pp respectively, or 12.0% and 8.9% of the base rate. Interestingly, the final column (7) shows 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 itself found.

6.2 Do Practitioners Already Know This?

Our procedure has identified two hypotheses that are new not only to the existing research literature, but also new 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? The practitioners might have learned the influence of these facial features from day-to-day experience. There have also been some studies showing a feature called “baby-faced” might influence decisions in mock-trial-type studies (Berry and Zebrowitz-McArthur, 1988).

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 pre-trial? Practitioners talked about judge misunderstandings of the law, people’s prior criminal records, and judge under-appreciation for the social contexts in which criminal records arise. Aside from the defendant’s race, nothing about the appearance of defendants was mentioned.

We then showed practitioners pairs of actual mugshots and asked them to guess which person is more likely to be detained by a judge (as we had done with M-turk and Prolific workers). This yields a sample of 360 detention forecasts. After seeing these mugshots we also asked practitioners 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 Figure XV.

The practitioner forecasts do indeed seem to be more accurate than those of “regular” study subjects. Column 5 of Table VII shows that defendants whom the practitioners predict will be detained are 29.2 pp 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 VI. 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 mugshot-based prediction is not so different from the adjusted R^2 from including both variables together in the same regression ($0.0165 + 0.0166 = 0.0331$ from columns 1 and 6, versus 0.0338 in column 7). We see something similar for the novel features of well-groomed and heavy-faced specifically as well.⁶⁴ The practitioners and the algorithm seem to be tapping into largely unrelated signal.

6.3 Exploring Causality

The over-arching goal of our procedure is to help generate ideas to feed into the causal-testing machinery that economists have developed. What might that sort of causal testing look like for the concrete empirical application we study here? 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.⁶⁵ In fact in our hometown of Chicago the court system has the defendant appear virtually but everyone else is in person, and the court system of its own volition has varied 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). Or 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 rather 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.

It is worth considering whether an empirically plausible hypothesis (out-of-sample correlation) is enough motivation to incur the costs of causal testing. It should be noted this is

⁶⁴The 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 column 2 of Table VII).

⁶⁵<https://www.nolo.com/covid-19/virtual-criminal-court-appearances-in-the-time-of-the-covid-19.html>.

probably at least as much empirical basis (and perhaps often more) than the informal standard economists currently use 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. Nonetheless given the substantial money and time costs of testing, in some cases researchers might want more than a simple correlation before proceeding to testing.

One (relatively) low-cost intermediate step is to measure and rule out some of the most obvious potential confounders, which become much easier to specify once specific hypotheses are in hand. 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, or is struggling with mental illness, or for that matter even more basically the defendant’s socio-economic status (SES). But as shown in a series of appendix tables, we find that when we have study subjects independently label the mugshots in our validation dataset and then control for them, the relationship between our novel hypotheses and both the algorithm’s predictions of the judge and the actual judge decisions persists.⁶⁶

A different (relatively) low-cost intermediate step researchers could take between the hypothesis-generation procedure and full-blown causal hypothesis testing would be something like a lab experiment. To illustrate, we carried out our own lab-style experiment with Prolific workers. We morphed synthetic mugshot images with respect to either well-groomed or heavy-faced, randomly assigned the structured variables relevant to each hypothetical case (current charge and prior record), 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. (See the appendix for details). 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, or because our appearance affects our own confidence, or because our appearance affects our own 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.

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

⁶⁶In 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. Appendix Tables A.X and A.XI show a qualitatively similar pattern of results for the defendant’s mental health and SES, which we measure by getting a separate sample of subjects to independently rate validation dataset mugshots for. We see qualitatively similar results when the dependent variable is the actual rather than predicted judge decision; see Appendix Tables A.XII, A.XIII and A.XIV.

the subjects also are more likely to recommend for detention those defendants who are less well-groomed or less heavy-faced. While of course only an actual field experiment could conclusively determine causality here, carrying out that type of field experiment might seem more worthwhile to an investigator in light of the lab experiment’s results.

7 Conclusion

We have presented here a new semi-automated procedure for hypothesis generation. We then 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.

Beyond the specific findings from our illustrative application, our empirical analysis here 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 subjects and get them to name what they see as the differences between morphed instances. Get other subjects 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 both clearly defined and measured for many cases. A study of, say, human creativity would be more challenging because it is not clear that can be measured (Said-Metwaly et al., 2017). A study of why US presidents use nuclear weapons during wartime would be challenging because of so few cases.

The second condition has to do with what input data are available to predict the behavior. Our procedure is likely to add only modest value in applications where we *only* have traditional structured variables. The reason is because those structured variables already make sense to people. Moreover the structured variables are usually already hypothesized to affect different behaviors, which is exactly why economists recommend asking questions about these variables 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.”

Moreover, 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. FOMC 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 such as 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 under-values a stock. Morphs of these time series might lead us to discover the kinds of price paths that produce over-reaction. 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 instances 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 accommodate a variety of unsupervised models. For example for text we can use other methods like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), or for time series we could use variational auto-encoders (VAEs) (Kingma and Welling, 2013).

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 hypotheses. 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 pre-specified 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 assembled by Leskovec et al. (2009). For generating ideas, creativity in experimental design useful for testing is replaced by 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. And 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 “pre-scientific” activity.

Tables

Table I: Summary statistics for Mecklenburg County NC Data, 2017-2020

	Train+Validation Set	Train Set	Validation Set	Untouched Lock-Box Data
Sample Size	32742	23138	9604	19009
Outcome				
Judge detains defendant	0.234	0.234	0.233	<i>Untouched</i>
Defendant re-arrested before trial	0.251	0.250	0.251	<i>Untouched</i>
Defendant Characteristics				
Age	31.793	31.859	31.631	<i>Untouched</i>
Male	0.787	0.789	0.782	<i>Untouched</i>
White	0.277	0.279	0.274	<i>Untouched</i>
Black	0.694	0.693	0.695	<i>Untouched</i>
Other	0.029	0.028	0.031	<i>Untouched</i>
Arrest Year				
2017	0.359	0.359	0.358	<i>Untouched</i>
2018	0.411	0.411	0.412	<i>Untouched</i>
2019	0.230	0.230	0.230	<i>Untouched</i>
Arrest Charge				
Violent	0.343	0.343	0.343	<i>Untouched</i>
Property	0.322	0.324	0.317	<i>Untouched</i>
Drug	0.205	0.204	0.207	<i>Untouched</i>
Gun	0.080	0.079	0.084	<i>Untouched</i>
Other	0.264	0.264	0.264	<i>Untouched</i>
Arrest Charge Severity				
Felony	0.424	0.422	0.428	<i>Untouched</i>
Non-Felony	0.576	0.578	0.572	<i>Untouched</i>
Defendant Prior Record				
Any Prior Conviction	0.462	0.463	0.458	<i>Untouched</i>
Prior Felony Conviction	0.332	0.334	0.328	<i>Untouched</i>
Prior Non-Felony Conviction	0.318	0.318	0.318	<i>Untouched</i>

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 untouched data set consists of data from the last 6 months of our study period (July 17, 2019, through January 17, 2019) plus a subset of cases through July 16, 2019, selected by randomly selecting arrestees. The remainder of the data set is then randomly assigned by arrestee to our training data set (used to build our algorithms) or our validation set (on which we report results in this paper draft). Once the paper is accepted, we will report final results for the untouched data set. For additional details of our data filters and partitioning procedures, see Table A.I. We define pre-trial release as being released on the defendant's own recognizance (ROR) or having been assigned and then posting cash bail requirements within three days of arrest. We define re-arrest as experiencing a new arrest before adjudication of the focal arrest, with detained defendants being assigned 0 values for 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. The multiple analyses of variance for the test of the joint null hypothesis that the difference in means across all variables is jointly zero has a p-value equal 0.4166. For pairwise p-values, please see the appendix Table A.II.

Table II: Is the algorithm rediscovering known facial features?

	<i>Dependent variable:</i>				
	Algo Judge Detain Prediction				
	(1)	(2)	(3)	(4)	(5)
Male	.1186*** (.0025)	.1179*** (.0025)	.1153*** (.0025)	.1138*** (.0025)	.1140*** (.0025)
Age		.0006*** (.0001)	.0006*** (.0001)	.0003*** (.0001)	.0003*** (.0001)
Black		.0029 (.0023)	-.0185*** (.0037)	-.0168*** (.0036)	-.0171*** (.0036)
Asian		-.0204* (.0115)	-.0232** (.0115)	-.0210* (.0114)	-.0216* (.0114)
Indigenous American		.0103 (.0241)	.0061 (.0240)	.0135 (.0238)	.0126 (.0238)
Skin-Tone			-.0441*** (.0059)	-.0411*** (.0058)	-.0417*** (.0058)
Attractiveness				-.0055*** (.0016)	-.0051*** (.0016)
Competence				-.0091*** (.0017)	-.0087*** (.0017)
Dominance				.0037*** (.0012)	.0030** (.0012)
Trustworthiness				-.0048*** (.0016)	-.0041** (.0016)
Human Guess					.0399*** (.0062)
Constant	.1595*** (.0022)	.1391*** (.0039)	.1771*** (.0064)	.2393*** (.0089)	.2173*** (.0095)
Observations	9,604	9,604	9,604	9,604	9,604
Adjusted R ²	.1954	.1992	.2038	.2195	.2228

Notes: The table above 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; 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 mugshot images from the Mecklenburg County, NC Sheriff's Office public website (see text). 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 III: Does algorithm predict judge behavior after controlling for known factors?

	<i>Dependent variable:</i>						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Algo Judge Detain Prediction	.6963*** (.0383)					.6262*** (.0433)	.6171*** (.0434)
Male		.1040*** (.0105)	.0978*** (.0106)		.0940*** (.0108)	.0228* (.0117)	.0244** (.0117)
Age		-.0008** (.0004)	-.0009** (.0004)		-.0013*** (.0004)	-.0015*** (.0004)	-.0015*** (.0004)
Black		-.0139 (.0098)	-.0651*** (.0156)		-.0618*** (.0156)	-.0513*** (.0154)	-.0521*** (.0154)
Asian		-.0753 (.0490)	-.0818* (.0490)		-.0754 (.0489)	-.0623 (.0484)	-.0638 (.0484)
Indigenous American		.0626 (.1024)	.0524 (.1023)		.0670 (.1021)	.0585 (.1011)	.0568 (.1010)
Skin-Tone			-.1059*** (.0251)		-.1004*** (.0251)	-.0747*** (.0249)	-.0762*** (.0249)
Attractiveness				-.0017 (.0063)	-.0053 (.0067)	-.0019 (.0067)	-.0011 (.0067)
Competence				-.0192*** (.0073)	-.0207*** (.0072)	-.0150** (.0072)	-.0144** (.0072)
Dominance				.0160*** (.0050)	.0095* (.0051)	.0071 (.0051)	.0057 (.0051)
Trustworthiness				-.0190*** (.0070)	-.0135* (.0071)	-.0105 (.0070)	-.0092 (.0070)
Human Guess							.0852*** (.0265)
Constant	.0576*** (.0106)	.1868*** (.0165)	.2780*** (.0272)	.3054*** (.0258)	.3928*** (.0381)	.2429*** (.0391)	.1981*** (.0415)
Naive-AUC	.625	.56	.571	.549	.586	.633	.635
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R ²	.0331	.0101	.0119	.0049	.0162	.0370	.0380

Notes: This table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables within the validation set described in Table 1. 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 (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 and unknown gender.

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

Table IV: Correlation between well-groomed (first novel feature) and algorithm's prediction

	<i>Dependent variable:</i>					
	Algo Judge Detain Prediction					
	(1)	(2)	(3)	(4)	(5)	(6)
Well-Groomed	-.0172*** (.0011)	-.0188*** (.0010)	-.0184*** (.0010)	-.0185*** (.0010)	-.0158*** (.0012)	-.0153*** (.0012)
Male		.1201*** (.0024)	.1192*** (.0024)	.1166*** (.0024)	.1153*** (.0025)	.1154*** (.0025)
Age			.0003*** (.0001)	.0002*** (.0001)	.0002*** (.0001)	.0002*** (.0001)
Black			.0050** (.0023)	-.0168*** (.0036)	-.0165*** (.0036)	-.0168*** (.0036)
Asian			-.0138 (.0113)	-.0165 (.0113)	-.0153 (.0113)	-.0160 (.0113)
Indigenous American			.0211 (.0237)	.0169 (.0236)	.0181 (.0236)	.0172 (.0236)
Skin-Tone				-.0449*** (.0058)	-.0437*** (.0058)	-.0440*** (.0058)
Attractiveness					.0006 (.0016)	.0008 (.0016)
Competence					-.0062*** (.0017)	-.0060*** (.0017)
Dominance					.0036*** (.0012)	.0031** (.0012)
Trustworthiness					-.0029* (.0016)	-.0024 (.0016)
Human Guess						.0339*** (.0062)
Constant	.3348*** (.0054)	.2486*** (.0051)	.2346*** (.0065)	.2736*** (.0082)	.2767*** (.0092)	.2568*** (.0099)
Observations	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R ²	.0247	.2249	.2262	.2310	.2337	.2361

Notes: This table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables, using data from the validation set of cases from Mecklenburg County, NC. 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 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 and unknown gender.

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

Table V: Correlation between heavy-faced (second novel feature) and algorithm's prediction

	<i>Dependent variable:</i>						
	Algo Judge Detain Prediction						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	-.0182*** (.0009)	-.0175*** (.0009)	-.0169*** (.0008)	-.0176*** (.0008)	-.0178*** (.0008)	-.0183*** (.0008)	-.0182*** (.0008)
Well-Groomed		-.0163*** (.0011)	-.0179*** (.0010)	-.0170*** (.0010)	-.0170*** (.0010)	-.0137*** (.0012)	-.0133*** (.0012)
Male			.1193*** (.0024)	.1180*** (.0024)	.1152*** (.0024)	.1127*** (.0024)	.1129*** (.0024)
Age				.0005*** (.0001)	.0005*** (.0001)	.0004*** (.0001)	.0004*** (.0001)
Black				.0057*** (.0022)	-.0179*** (.0035)	-.0181*** (.0035)	-.0183*** (.0035)
Asian				-.0115 (.0111)	-.0145 (.0110)	-.0134 (.0110)	-.0140 (.0110)
Indigenous American				.0078 (.0232)	.0030 (.0231)	.0046 (.0230)	.0039 (.0230)
Skin-Tone					-.0488*** (.0057)	-.0469*** (.0057)	-.0472*** (.0056)
Attractiveness						-.0035** (.0016)	-.0034** (.0016)
Competence						-.0062*** (.0016)	-.0061*** (.0016)
Dominance						.0063*** (.0012)	.0058*** (.0012)
Trustworthiness						-.0004 (.0016)	.00003 (.0016)
Human Guess							.0286*** (.0060)
Constant	.3485*** (.0050)	.4230*** (.0070)	.3340*** (.0065)	.3133*** (.0073)	.3568*** (.0089)	.3597*** (.0098)	.3423*** (.0104)
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R ²	.0384	.0603	.2579	.2613	.2669	.2711	.2727

Notes: This table shows the results of estimating a linear probability specification regressing algorithmic prediction of judges' detain decision against different explanatory variables, using data from the validation set of cases from Mecklenburg County, NC. Algorithmic predictions of judges' decisions come from applying 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 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 and unknown gender.

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

Table VI: Do well-groomed and heavy-faced (first and second novel features) correlate with judge decisions?

	<i>Dependent variable:</i>						
	Judge Detain Decision						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Heavy-Faced	-.0234*** (.0036)		-.0226*** (.0036)	-.0223*** (.0036)		-.0218*** (.0037)	-.0111*** (.0037)
Well-Groomed		-.0198*** (.0043)	-.0185*** (.0043)		-.0124** (.0051)	-.0100* (.0051)	-.0022 (.0051)
Algo Judge Detain Prediction							.5842*** (.0449)
Male				.0918*** (.0107)	.0959*** (.0108)	.0928*** (.0108)	.0269** (.0118)
Age				-.0011*** (.0004)	-.0013*** (.0004)	-.0012*** (.0004)	-.0014*** (.0004)
Black				-.0645*** (.0156)	-.0624*** (.0156)	-.0643*** (.0156)	-.0535*** (.0154)
Asian				-.0737 (.0488)	-.0726 (.0489)	-.0701 (.0488)	-.0620 (.0484)
Indigenous American				.0490 (.1019)	.0683 (.1021)	.0524 (.1019)	.0501 (.1010)
Skin-Tone				-.1062*** (.0250)	-.1038*** (.0251)	-.1076*** (.0250)	-.0801*** (.0249)
Attractiveness				-.0084 (.0067)	.0004 (.0070)	-.0045 (.0070)	-.0025 (.0070)
Competence				-.0194*** (.0072)	-.0175** (.0073)	-.0176** (.0073)	-.0141* (.0072)
Dominance				.0109** (.0052)	.0076 (.0051)	.0108** (.0052)	.0075 (.0051)
Trustworthiness				-.0085 (.0071)	-.0104 (.0071)	-.0075 (.0071)	-.0075 (.0070)
Human Guess				.1023*** (.0267)	.1049*** (.0268)	.0986*** (.0268)	.0819*** (.0266)
Constant	.3569*** (.0196)	.3280*** (.0209)	.4418*** (.0276)	.4436*** (.0446)	.3642*** (.0429)	.4665*** (.0462)	.2666*** (.0483)
Naive-AUC	.544	.531	.553	.601	.592	.601	.637
Observations	9,604	9,604	9,604	9,604	9,604	9,604	9,604
Adjusted R ²	.0042	.0021	.0061	.0215	.0183	.0218	.0387

Notes: This table reports the results of estimating a linear probability specification of judges' detain decisions against different explanatory variables within 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 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. Regression specifications also include indicators for unknown race and unknown gender.

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

Table VII: Results from criminal justice practitioner sample

	<i>Dependent variable:</i>							
	Judge Detain Decision							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Criminal Justice Practitioner Guess	.4172*** (.1576)		.3635** (.1592)		.2924* (.1593)		.4244*** (.1562)	.3395** (.1567)
Algo Judge Detain Prediction						.6201*** (.2335)	.6307*** (.2315)	.7555*** (.2717)
Well-Groomed		-.0455* (.0261)	-.0362 (.0263)	-.0273 (.0305)	-.0206 (.0306)			
Heavy-Faced		-.0394* (.0217)	-.0363* (.0216)	-.0411* (.0217)	-.0387* (.0217)			
Male				-.0696 (.0655)	-.0680 (.0653)			-.1579** (.0725)
Age				-.0036 (.0029)	-.0035 (.0029)			-.0032 (.0028)
Black				-.1683* (.0934)	-.1706* (.0931)			-.1454 (.0926)
Skin-Tone				-.3901** (.1568)	-.3895** (.1562)			-.3192** (.1562)
Attractiveness				-.0062 (.0448)	-.0090 (.0447)			.0049 (.0432)
Competence				.0021 (.0441)	.0039 (.0440)			.0005 (.0434)
Dominance				.0512* (.0307)	.0475 (.0307)			.0334 (.0304)
Trustworthiness				-.1113** (.0446)	-.1031** (.0447)			-.1145** (.0443)
Constant	.2855*** (.0831)	.9205*** (.1662)	.6778*** (.1965)	1.4446*** (.2728)	1.2442*** (.2929)	.3377*** (.0646)	.1226 (.1018)	.7930*** (.2679)
Naive-AUC	.572	.577	.602	.643	.653	.576	.607	.661
Observations	360	360	360	360	360	360	360	360
Adjusted R ²	.0165	.0131	.0246	.0384	.0449	.0166	.0338	.0582

Notes: This table shows the results of estimating judges' detain decision 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 decision comes from showing 15 different public defenders and legal aid society members actual mugshot images of defendants and asking them to report which defendant they believe the judge would be more likely to detain. The pairs are selected in such way 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 mugshot images (see text). Regression specifications also include indicators for unknown race and unknown gender.

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

Figures



Figure I: Illustrative facial images

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

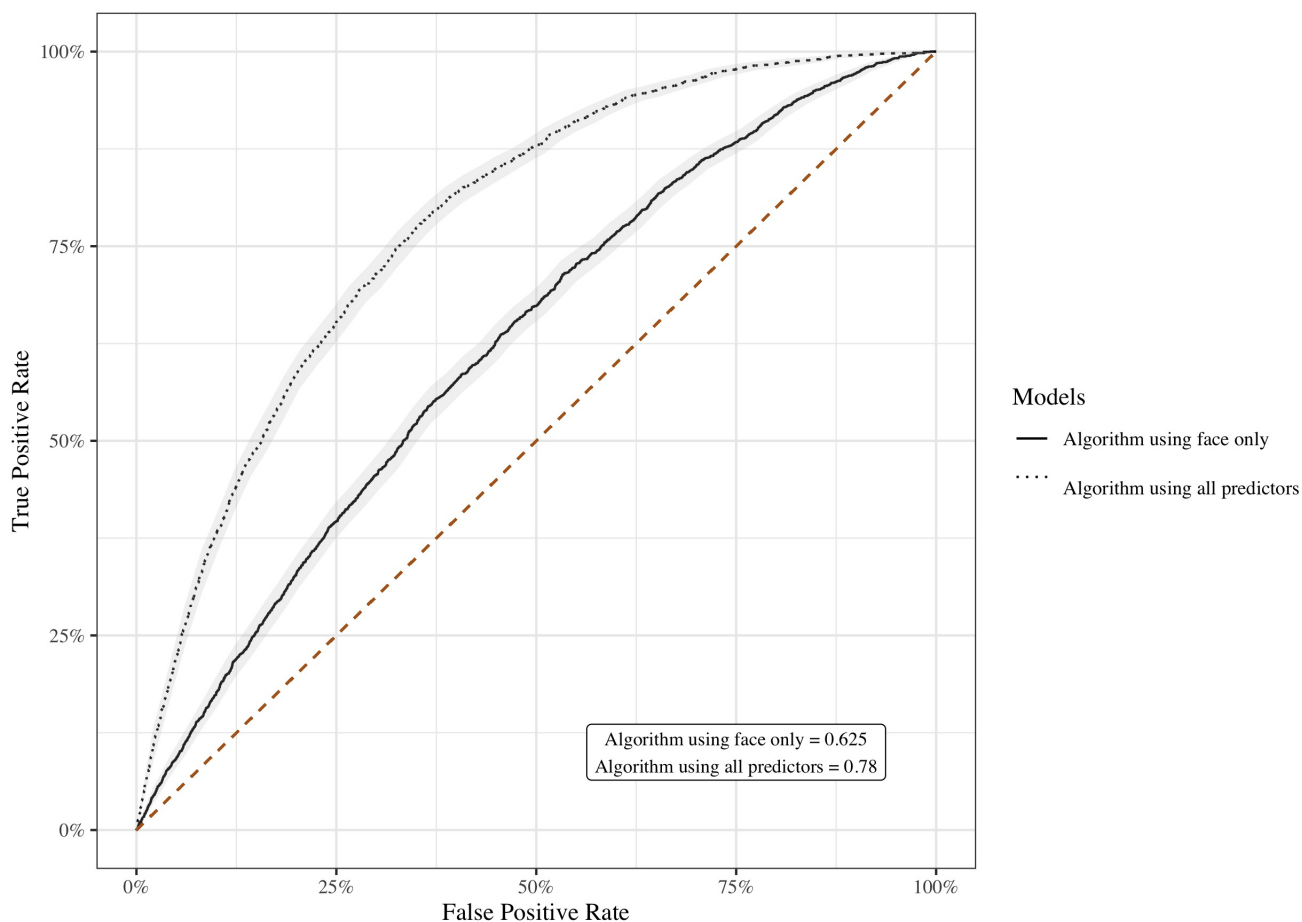


Figure II: Accuracy of algorithmic models of judge decisions

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.

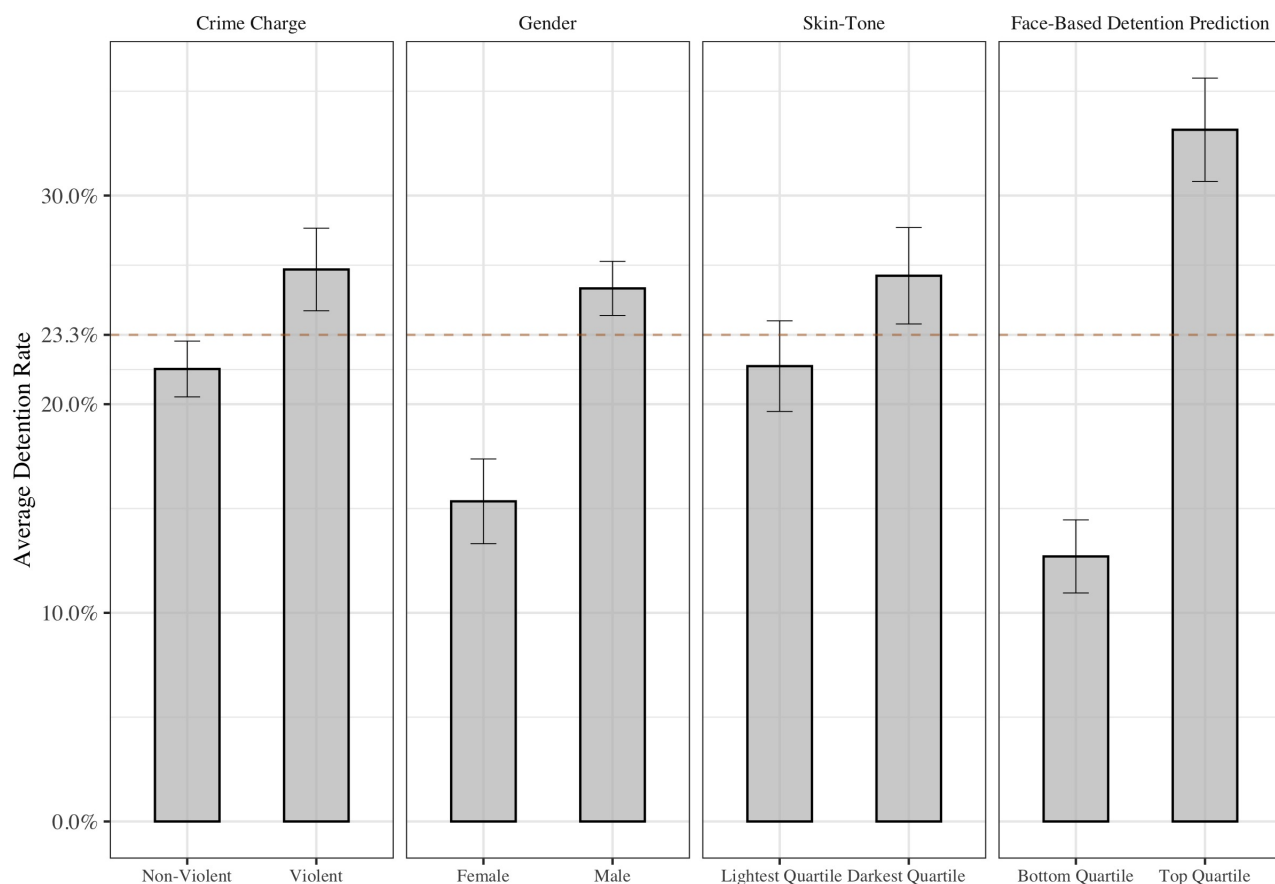
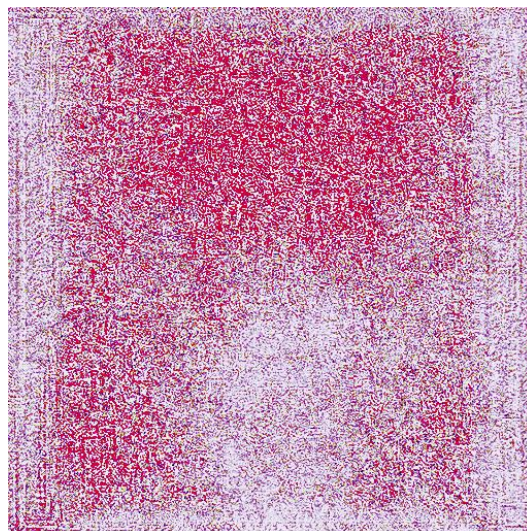


Figure III: Relationship between detention rates and defendant characteristics

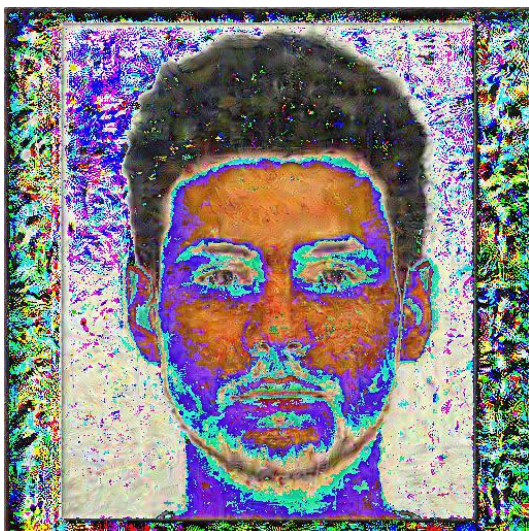
Notes: The figure above shows the average validation set detention rates for defendants by different defendant characteristics: crime charge is violent vs. non-violent (first panel), defendant is male versus female (second panel), defendant is in the lightest (Q4) versus darkest (Q1) skin tone shade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



(a) Initial face



(b) Saliency map



(c) Naive age-morphed image



(d) Morphs from our procedure

Figure IV: Candidate algorithm-human communication vehicles for a known facial feature: Age

Notes: The first panel shows a randomly selected point in the GAN latent space for a non-Hispanic white male defendant. The second panel shows a saliency map that highlights the pixels that are most important for an algorithmic model that predicts the defendant's age from the mugshot image. The third panel 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 text). The final panel shows the image morphed to the maximum age using our actual preferred morphing procedure.

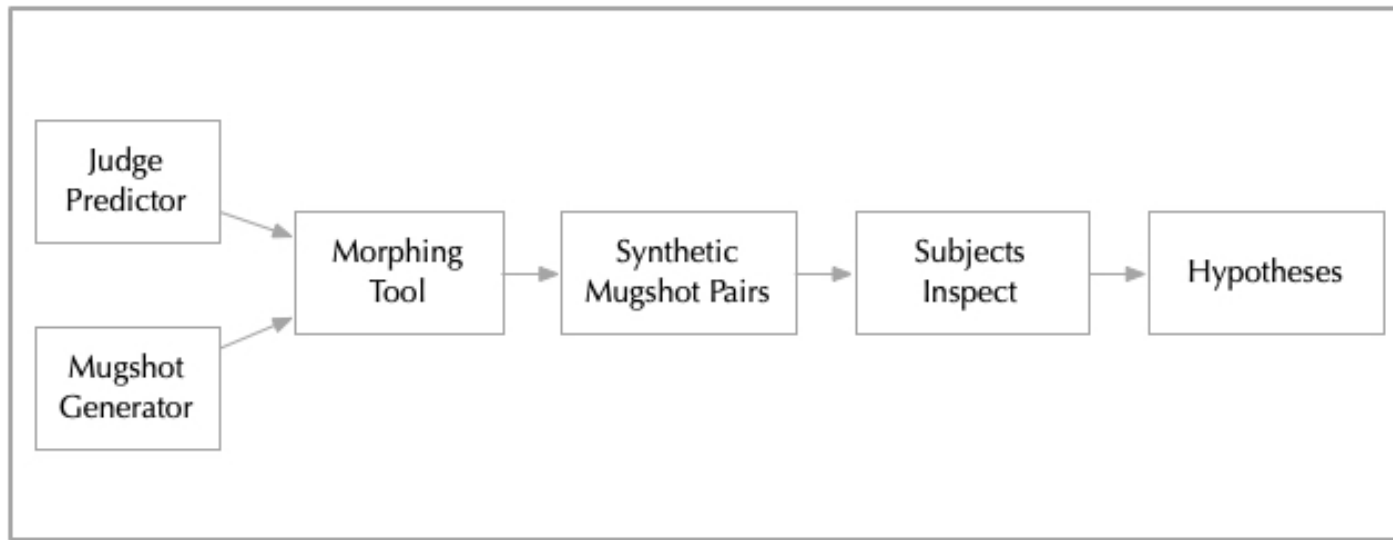
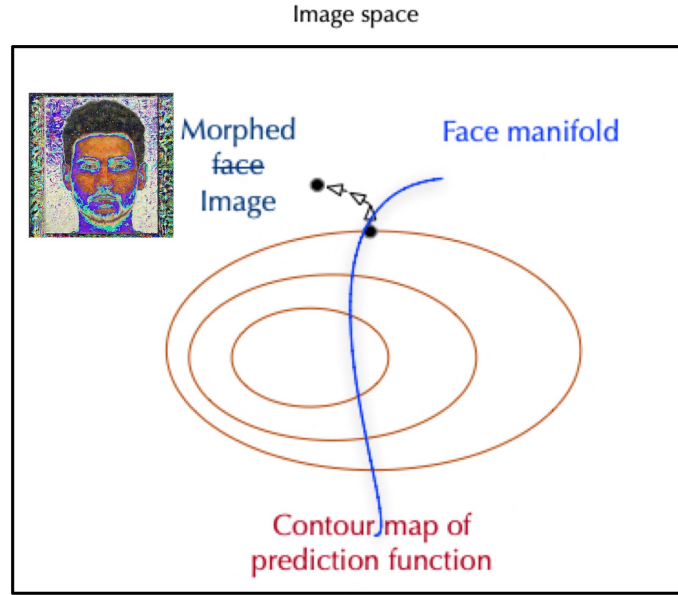
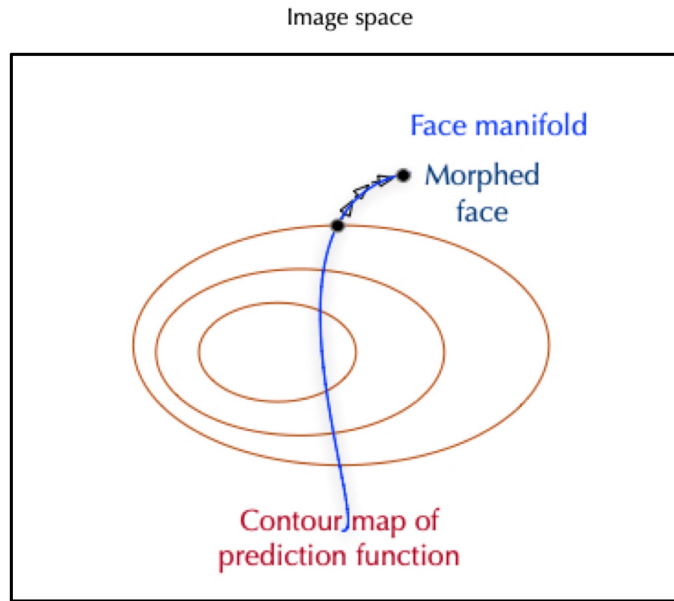


Figure V: Hypothesis generation pipeline

Notes: The above diagram illustrate all the algorithmic components in our procedure by presenting a full pipeline for algorithmic interpretation.



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



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

Figure VI: Morphing images for detention risk on and off the face manifold

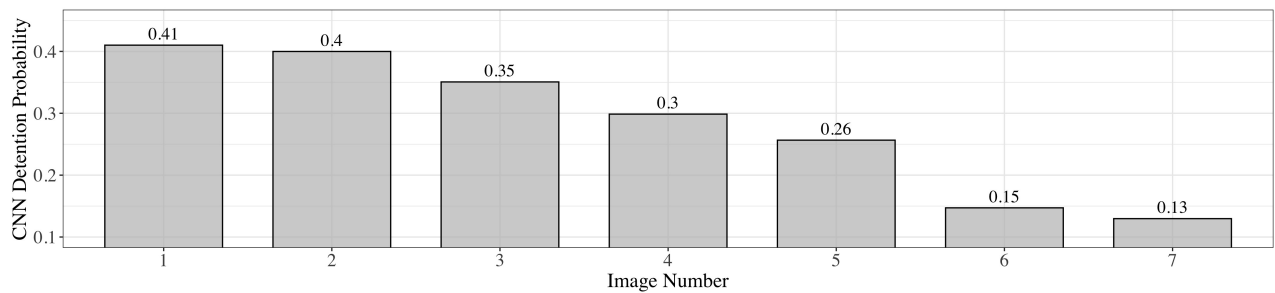
Notes: The exhibits above show the difference between an unconstrained (naive) morphing procedure and our preferred new morphing approach. In both panels, the background represents the image space (set of all possible pixel values) and the blue line represents the set of all pixel values that correspond 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 non-face 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).



(a) Side-by-side mugshot 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 VII: Illustration of morphed faces along detention gradient

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 new morphing procedure to increase the predicted detention risk of the image to 0.41 (at left) or reduce the predicted detention risk down to 0.13 (at right). The overall average detention rate in the validation data set 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 VIII: Examples of morphing along the gradients of face-based detention predictor

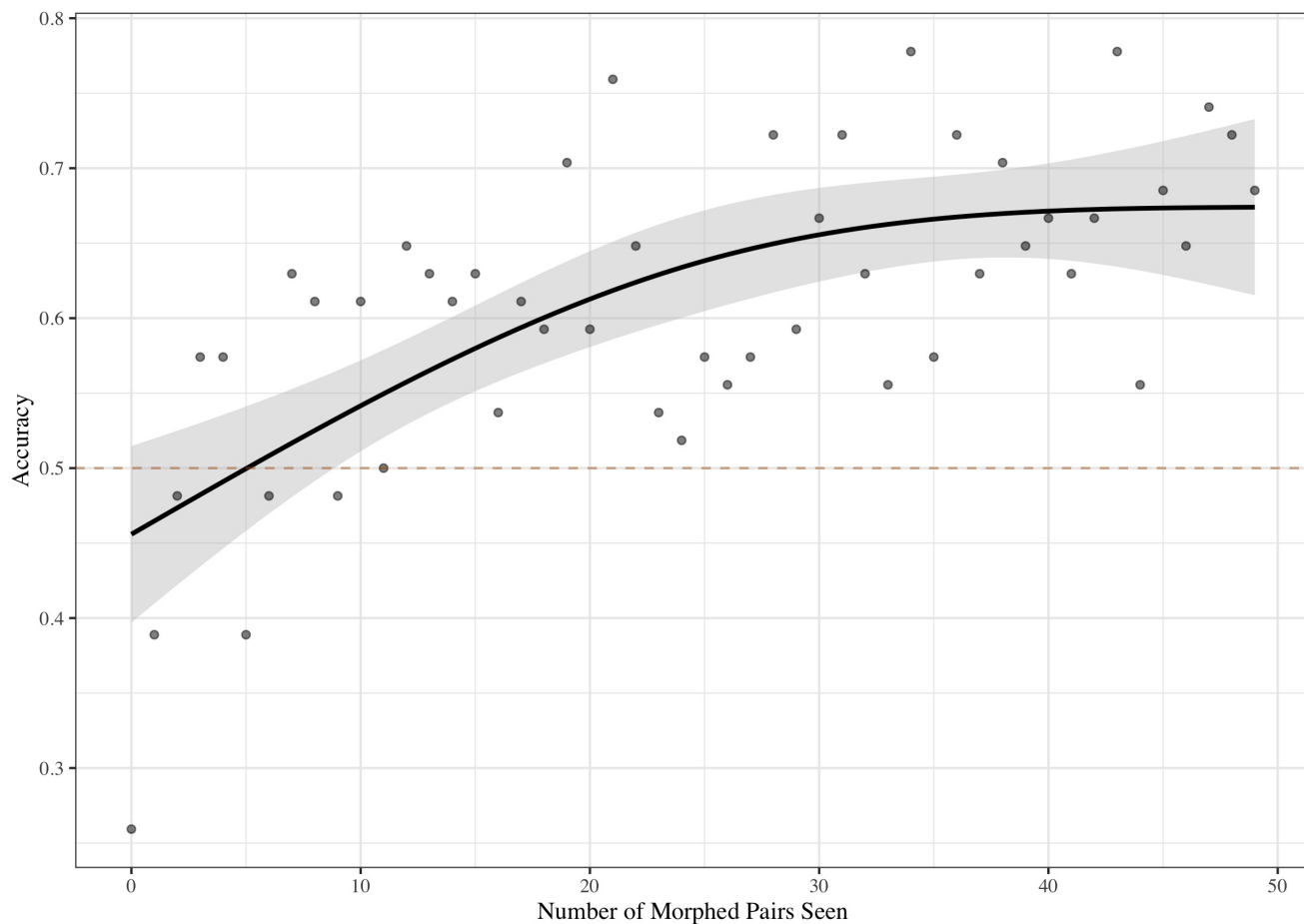


Figure IX: Subject performance guessing relative detention risk across morphed image pairs as a function of number of images seen

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

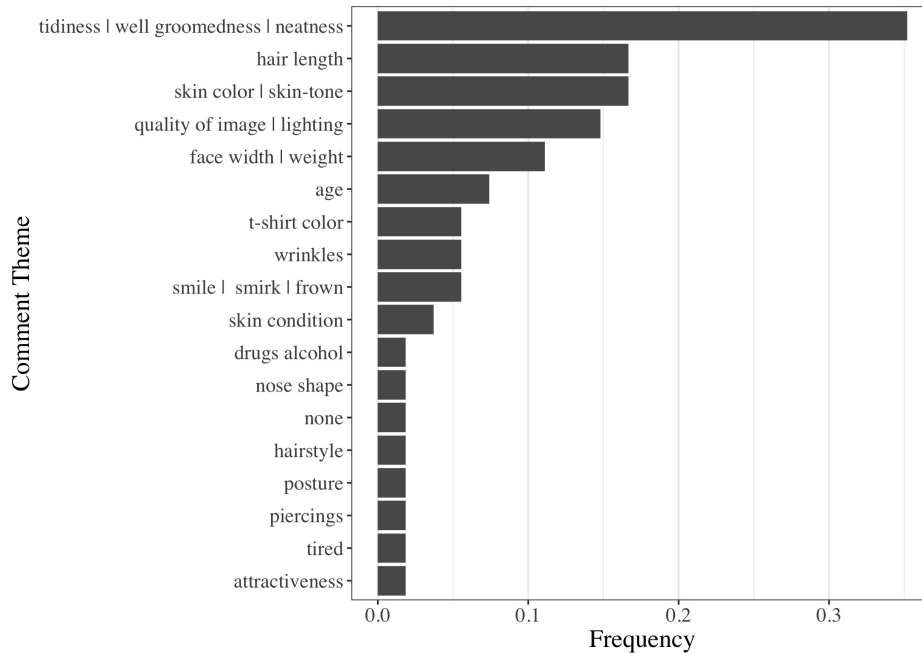


Figure X: Subject reports of what they see between detention-risk-morphed image pairs

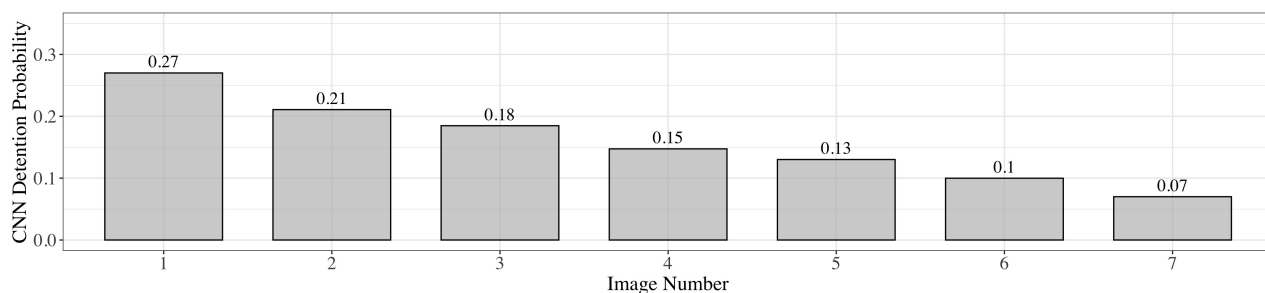
Notes: The top panel shows a word cloud of subject reports about what they see as the key difference between image pairs where one is a randomly selected point in the GAN latent space and the other is morphed in the direction of a higher predicted detention risk. Words are approximately proportionately sized to the frequency of subject mentions. The bottom panel shows the frequency of semantic groupings of those open-ended subject reports (see text for additional details).



(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



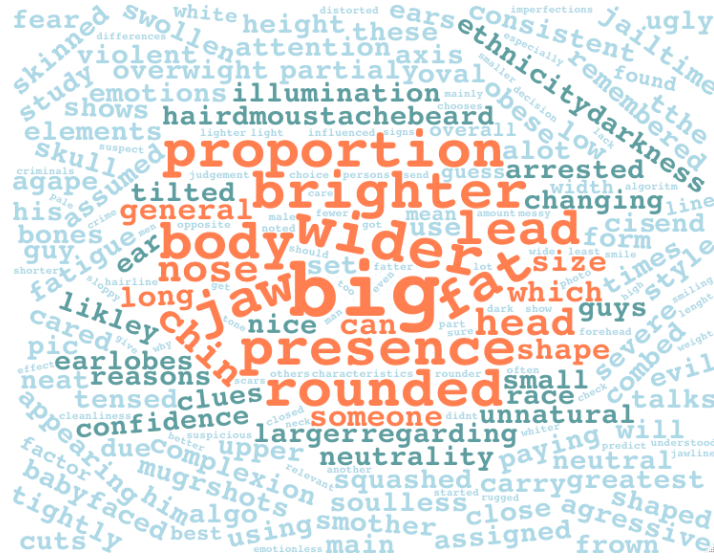
(c) Detention-probabilities for images in panel (b)

Figure XI: Illustration of morphed faces along orthogonal gradients of detention predictor

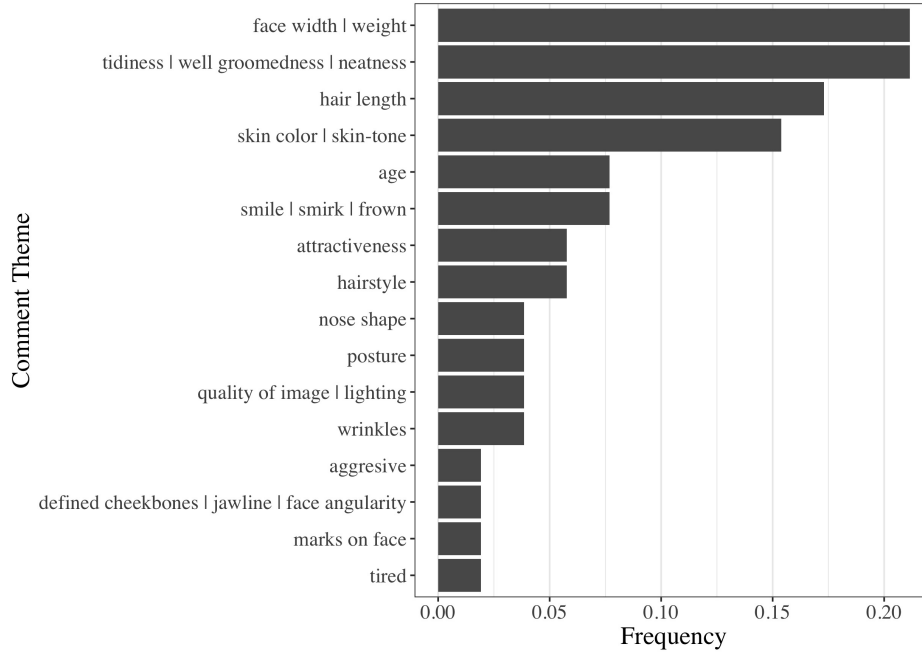
Notes: The top panel shows the result of selecting a random point on the GAN latent face space for a white Hispanic male defendant, then using our orthogonal morphing procedure to increase the predicted detention risk of the image to 0.27 (at left) or reduce the predicted detention risk down to 0.07 (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 XII: 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 XIII: Subject reports of what they see between detention-risk-morphed image pairs, orthogonalized to first novel feature discovered (well-groomed)

Notes: 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 point in the 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 text). The bottom panel shows the frequency of semantic groupings of these open-ended subject reports (see text for additional details).

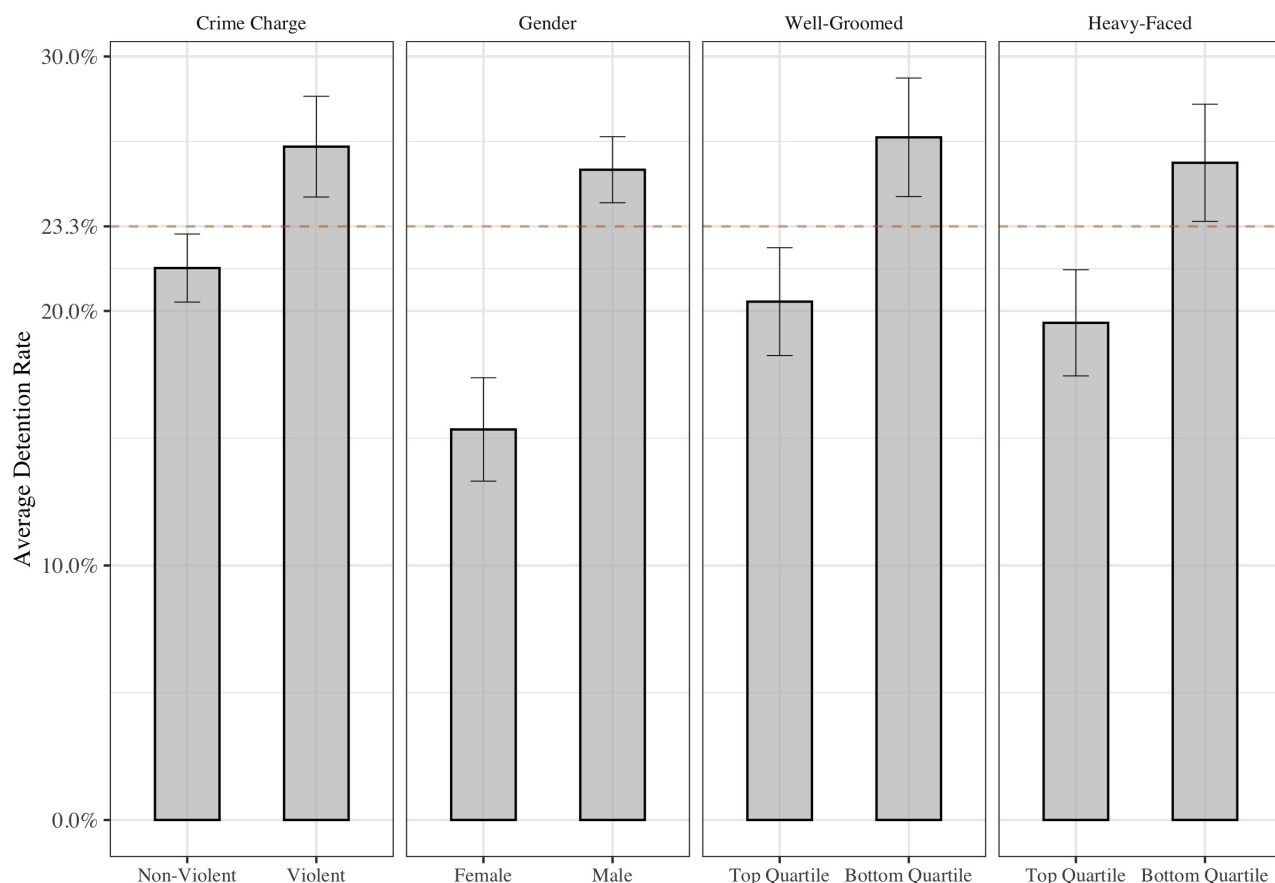
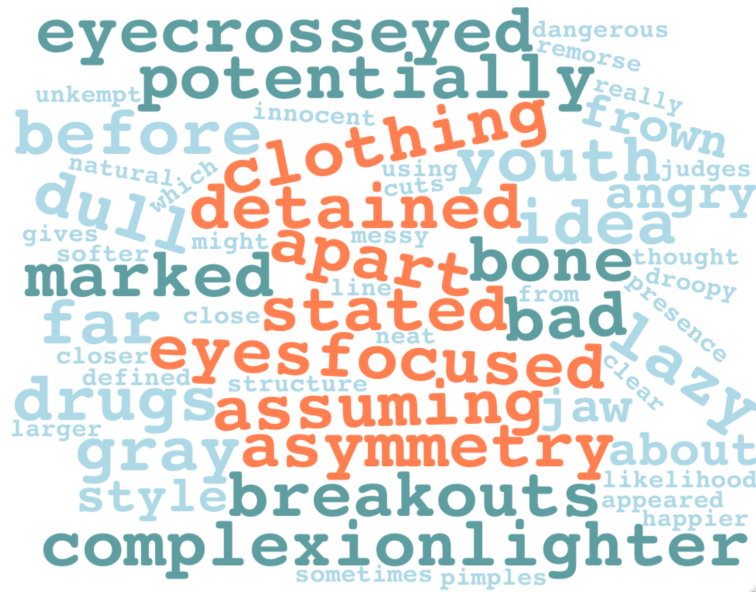
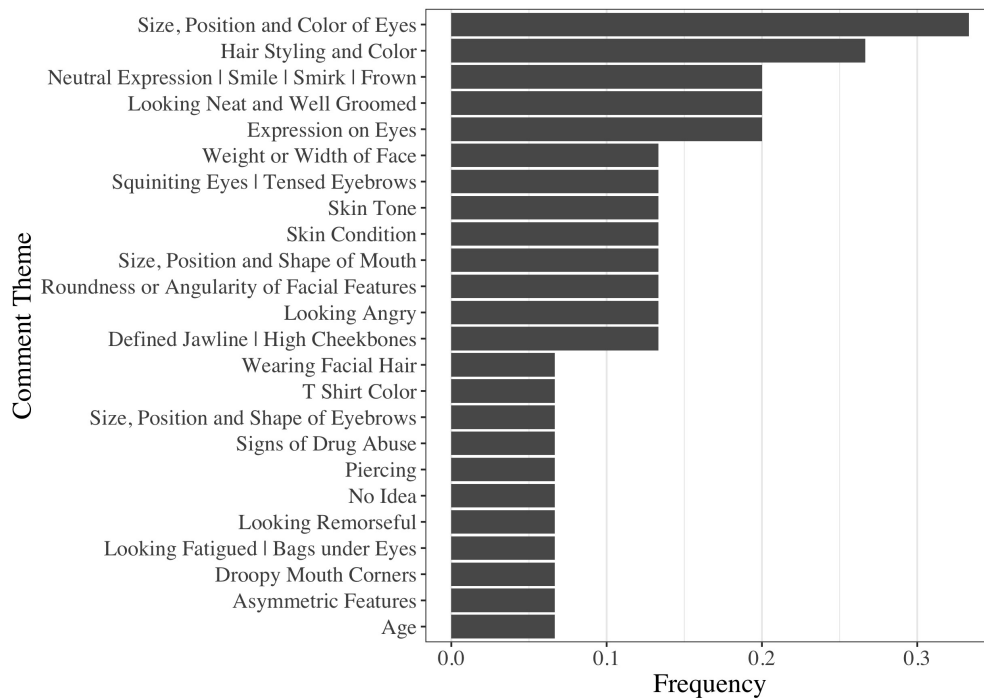


Figure XIV: Relative magnitude of the algorithm's discoveries on detention

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 XV: 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).

A Appendix A: Conceptual Framework

A.1 A Simple Framework for Discovery

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

A.2 Setup

We assume the overall goal is to understand how some outcome (y) depends on some input (x). Simply finding any function from x to y is not enough. Instead, for us, understanding requires that we uncover “interpretable” functions that relate the two. The goal of hypothesis generation is to uncover candidate interpretable functions. Testing—which we do not model—happens next: it verifies whether a given candidate function in fact explains the relationship between y and x . The hypotheses are to be generated from some data, which we assume consist of (y, x) pairs that are from a distribution $\mathcal{D}$. We will also assume that $x \in \mathbb{R}^k$ is a k -dimensional real vector, $y \in \{0, 1\}$ is some binary outcome, and that there is some function $f^* : \mathbb{R}^k \rightarrow \{0, 1\}$ such that $y = f^*(x)$ for all (x, y) sampled from $\mathcal{D}$.⁶⁷

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

In this framework, we are not interested in the estimates produced by f^* , but rather in uncovering “interpretable” functions that explain how x relates to y . For example, skin color and attractiveness (as judged by a certain group) are interpretable functions that can be calculated for a given image x . In the diary example, interpretable functions might include a measure of whether the text is “sad”, or if it “discusses self more than others,” or “discusses emotions”; all of these are interpretable functions of the input vector which have some meaning to people. But the converse is not true: most mathematical functions of images are not interpretable, and have no meaning to us. With this in mind, let $\mathcal{H}$ be the set of all possible hypotheses—the set of every function from input data x to the chosen outcome y . We assume there is also some set $\mathcal{I} \subseteq \mathcal{H}$ of *interpretable* functions, which are the functions that admit some human-understandable description.⁶⁸ We also assume for a given

⁶⁷Relaxing the simplifying assumptions that y is binary and that there is no noise in the relationship between x and y does not substantively change the setup or the substance of our results.

⁶⁸Note the set $\mathcal{I}$ is not static: fundamental innovations such as the discovery of calculus or probability can

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

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

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

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

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

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

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

It is worth pointing out what makes this task hard (or easy). Suppose we simply built a model $m(x)$ that predicts y . For that model itself to yield an interpretable hypothesis, its parameters must be interpretable. That can happen in some simple cases. For example, if we had a data set where each dimension of x was interpretable (such as individual structured variables in a tabular dataset) and we used a predictor such as OLS (or LASSO), then we could just read the hypotheses from the non-zero coefficients: which variables are significant? Even in that case, interpretation is challenging because machine learning tools, built to generate accurate predictions, yield coefficients that can be quite unstable and change with small perturbations in the data (Mullainathan and Spiess, 2017).⁶⁹ And often interpretation is much less straightforward than that. If x is an image, text or time series, the estimated models (such as convolutional networks) are defined on granular inputs and have no particular meaning: if we knew the algorithm weighted a particular pixel, what have we actually learned? In these cases, the estimated model m is itself not interpretable, nor is it readily apparent how we might decompose m into interpretable sub-components in the same way that we have decomposed f^* in (1). Our focus is on these contexts where algorithms, as “black-box” models, are not readily interpreted.

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

⁶⁹The intuition here is quite straightforward: If two predictor variables are highly correlated, the weight that the algorithm puts on one versus the other can change from one draw of the data to the next depending on the idiosyncratic noise in the training dataset, but since the variables are highly correlated the predicted outcome values themselves (hence predictive accuracy) can be quite stable.

A.3 The Discovery Problem

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

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

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

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

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

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

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

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

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

For all three criteria, probabilities and expectations are taken over new draws from the original data-generating process $\mathcal{D}$.⁷⁰

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

⁷⁰As a result, the performance of a given procedure can depend on $\mathcal{D}$, which is intuitive: some procedures may do better for some data-generating processes than others. Additionally, We could have included another criterion here: *novelty*. Because some of the hypotheses $\phi_i^*(x)$ may already be discovered, we may wish our hypothesis generation procedure to generate a *new*, previously undiscovered hypothesis. We exclude this here because we could imagine the procedure being applied to $y - \phi'(x)$ where ϕ' is the already known factor that influences y .

replicability will follow. But if there are many plausible hypotheses, then this need not be the case. Third, crucially, note that *veracity* is not included as a criterion; that is, h need not necessarily be true. We do not require that $h \in \Phi^*$, nor do we require that changing $h(x)$ would change y .⁷¹ Requiring veracity sets entirely too high a bar. Finally, both plausibility and comprehensibility can just as easily be defined as properties of the individual hypothesis generated by a procedure (we will apply these definitions below in our empirical work). In contrast, replicability can only be defined as a property of the overall procedure.

A.4 Human hypothesis generation

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

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

Cognitive Constraints: *Uniformly sampling (without replacement) n matched pairs to form a set E .⁷²*

Inspiration: *Picking one hypothesis at random from $\mathcal{I}_E$, the set of interpretable hypotheses consistent with E .*

Advantageous prior: *Possibly picking true hypotheses with higher probability. With probability α , the hypothesis is chosen from Φ^* , the set of true hypotheses. With probability $1 - \alpha$, the hypothesis is picked uniformly at random from $\mathcal{I}_E \setminus \Phi^*$, the set of untrue interpretable hypotheses consistent with E .*

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

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

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

⁷¹Specifically, if we changed only the part of x that affected $h(x)$ and held the rest constant, y should change. Importantly, note that veracity (unlike plausibility) does not depend on the data-generating process: it is a feature of the true function, not the specific way we draw data.

⁷²Notice in our rendition, for simplicity, we have not allowed for intrinsic biases in what humans might notice, such as confirmatory or categorical biases that might lead them to systemically notice certain relationships that are not there. Such biases would worsen both human hypothesis generation and our suggested procedure.

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

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

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

A.5 Algorithmic hypothesis generation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Left to their own devices, people seek to identify *any* differences across data points that differ in y . Because they only have one particular data set, that approach is prone to over fitting: differences that may hold in that particular sample but not others. But with our morphing procedure, humans are now looking for differences in $m(x)$, and we know $m(x)$ is a reliable out-of-sample predictor. By way of intuition, this is why we are usually better off interpreting a regression than individual data points that go into the regression. Although there is no noise in the setup we use here, the same basic intuition carries through.
- The matching of nearest neighbors reduces the curse of dimensionality. Pairs (x^-, x^+) will now have fewer plausible candidates for what may be different between them exactly because we have ensured they differ as little as possible.⁷³

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

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

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

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

⁷³An analogy with OLS provides an easy intuition for why this helps: a regression is in effect a way to calculate in-sample correlations, and the standard errors tell us whether those correlations are real or due to sampling noise. Adding controls lowers standard errors because lowering the residual noise makes it more likely that in-sample correlations are more likely to hold out-of-sample. That is the same effect here: by matching on $m(x)$ instead of y , we are reducing residual noise and increasing the chance that noticed patterns are genuine ones.

- (ii) *The problem is complex enough. That is, the relative number of comprehensible hypotheses that are true is sufficiently small: $\frac{|\Phi^*|}{|\mathcal{I} \setminus \Phi^*|} < C_\Phi$, and*
- (iii) *Implausible hypotheses are consistent with matching morph pairs sufficiently rarely: $\mathbb{P}_\mathcal{D}(h \in \mathcal{I}_{e_m}) < C_m$ for any hypothesis $h \in \mathcal{I} \setminus \Phi^*$ and any random matching morph pair e_m ,*

then

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

(See Section ?? for a sketch proof of Proposition 1.)

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

Finally, it is worth noting a few additional advantages of algorithmic generation that are not highlighted by this proposition. First, though it is not explicit here, given a set of *known* hypotheses, we can orthogonalize with respect to those dimensions to ensure that the algorithm is producing something novel.⁷⁴ Second, other methods of producing hypotheses (observation, conversation, introspection) may produce theories that are hard to measure in data. By construction, our procedure only produces hypotheses that are measurable.

A.6 A proof of Proposition 1

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

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

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

⁷⁴In our particular application, we do not do this, in part because we are curious to explicitly examine algorithms’ capacity to rediscover known hypotheses.

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

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

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

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

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

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

which implies that

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

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

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

which implies that

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

and hence that

$$\alpha < (1 - \beta_m)^K. \quad (3)$$

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

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

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

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

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

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

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

Conversely, for the morphing discovery process, we see that

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

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

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

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

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

Using complementarity and the assumption of equiprobability, we know that

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

Hence,

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

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

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

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

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

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

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

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

which proves the first claim of the proposition.

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

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

Hence,

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

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

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

Then (9) becomes

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

Then we can use (9) to see that

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

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

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

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

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

which proves the second claim of the proposition, and completes the proof. $\square$

B Appendix B: Data and Institutional Details

B.1 Pre-trial detention decisions

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

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

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

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

In the specific jurisdiction from which we have obtained data here, Mecklenburg County, North Carolina, the very first hearing for the defendant is overseen not by a judge, but by a “magistrate” (who is like a judge, but is not elected). Defendants not released by the magistrate are booked into jail and see a judge the next day (Redcross et al., 2019). Starting in 2014, judges were given access to a pre-trial risk prediction tool developed by the Arnold Foundation called the Public Safety Assessment (Redcross et al., 2019). The PSA gives judges predictions from a logistic regression for three separate outcomes: (1) risk of failure to appear (FTA) in court at a required future hearing; (2) risk of any new criminal activity (NCA); and (3) risk of any new violent criminal activity (NVCA). The PSA makes these predictions using factors like age, current charge, and prior record.⁷⁵ Because defendants can only be detained if the magistrate and judge agree on detention, and because the magistrate’s decision is made in the shadow of the judge, and because (more pragmatically) the data we have do not separately identify the magistrate’s decision from that of the judge, we follow Redcross et al. (2019) and combine both decisions into a single detain-versus-release outcome.

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

⁷⁵See <https://advancingpretrial.org/psa/factors/>.

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

B.2 Mecklenburg County criminal justice data

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

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

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

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

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

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

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

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

From the raw data we construct features corresponding to:

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

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

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

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

C Appendix C: Methods

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

C.1 Predicting judges’ decisions and defendant risk

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

As noted in the text, we build separate types of models for the structured data (gradient boosted decision trees) and unstructured data (convolutional neural networks, or CNNs). For our models that rely on both structured and unstructured data, we use a stacking procedure that forms new predictions that are weighted averages of the structured data predictor and unstructured data predictor, with the data used to select the weight. Since we are using standard machine learning methods at this stage of our analysis, we focus our discussion here on high-level descriptions.⁷⁶

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

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

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

The logic behind the CNN method is perhaps easiest to see by considering its alternatives. To an algorithm, a 512×512 black-and-white image is essentially just 262,144 pixel values.⁷⁷ 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

⁷⁶For excellent overviews of decision trees and gradient boosting methods at various levels of technical detail, see for example Bishop and Nasrabadi (2006), Breiman (2001), Breiman et al. (2017), Freund et al. (1999), Hastie et al. (2009), and James et al. (2013). Examples of excellent discussions of deep-learning methods at various levels of technical complexity include Yegnanarayana (2009), LeCun et al. (2010), Krizhevsky et al. (2012), LeCun et al. (2015), Nielsen (2015), Rawat and Wang (2017), and Gurney (2018).

⁷⁷For a color image, there are three times as many values, since pixels have red, blue and green shadings.

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

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

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

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

The image data are fed into a 50-layer residual network (“resnet50”) that consists of 4 convolutional blocks and 2048 output neurons, using a gentle decay learning rate schedule (see He et al. (2016)). Because the more basic features of images are not specific to the types of images being analyzed, we can improve performance of this network by pre-training it on a separate set of images. The resnet50 we use here was pre-trained on ImageNet data⁷⁸ 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

⁷⁸<https://www.image-net.org/>

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

C.2 Alternative methods for algorithmic interpretability

The problem we face—understanding what our algorithm sees in the face—has emerged as a central challenge in machine learning research. A variety of techniques have been developed for interpreting or explaining how machine learning algorithms form their predictions (see Marcinkevičs and Vogt (2020) for a recent review).⁷⁹ 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 explainability methods can only provide explanations using measured high-level features. For example, Li et al. (2018), Ghorbani et al. (2019), Zhang et al. (2018), and Chen et al. (2019) among others develop interpretability tools that highlight not individual pixels that are important for classification, as in saliency maps, but higher-level *concepts* or prototypical parts within these images, such as wheels helping classify the presence of a van in an image. But all these approaches require the explanatory features to already be coded: the data must contain for each image, for example, information on whether “wheel” was present or not.⁸⁰ In these examples, the goal is typically not discovery but instead either to explain the model to people to aid in decisions, sometimes as required by explanation (Wachter et al., 2018), or to assess the robustness of models, such as whether a breast cancer detector is looking in the right place (Bai et al., 2021). Moreover, since the potential explanations are already in the data set, one could go further: rather than building a black-box model and explaining it, build one that is explainable to begin with.⁸¹ 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,

⁷⁹For 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.

⁸⁰In 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.

⁸¹See, for example, Holte (1993), Rudin et al. (2010), Freitas (2014), Letham et al. (2015), Angelino et al. (2018), Jung et al. (2017), Chen and Rudin (2018), Ustun and Rudin (2019), Rudin (2019), and the references therein.

the goal would be to generate an old face, and this is done when age is measured in the data during training.⁸²

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.⁸³ In some sense, it is akin to the problem of hypothesis generation: what features should we annotate?

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

C.3 Generative Adversarial Networks

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

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

⁸²One 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.

⁸³A recent ambitious paper has tried to tackle this problem. Peterson et al. (2022) collected millions of labels on hundreds of facial features and then created a predictive model of them for synthetic faces. The challenge, however, is that this model is built on synthetic faces, whereas we would need such a model for actual images (mugshots). Deep learning models are known to not transfer across distributions. In fact, when we attempt to use the results of this paper, we find our mugshots do not map into these synthetic faces in any meaningful way. The failure is a reminder that while humans tend to think of “faces” in the abstract, algorithms model very specific distributions of pixel combinations. It is why we must build our own generative model of mugshots rather than use extremely well-developed generative models of “faces.”

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

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

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

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

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

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

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

One common measure for assessing a GAN’s quality is the Frechet inception distance (FID) (Heusel et al., 2017), which is a measure of the difference between the distribution of GAN-generated images relative to the original images used to train the GAN.⁸⁴ 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 2.84.⁸⁵ We likely do better because the space of mugshots is a smaller, less rich space than the space of faces in the Flickr dataset.

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

⁸⁴Calculation 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.

⁸⁵As noted above, to avoid stereotyping in discussions of crime and criminal justice, we illustrate the key ideas in our paper using images just for non-Hispanic white males. So the GAN performance statistics we report here are from a StyleGAN2 trained just on males in our mugshot data set, which as shown in Table I accounts for the large majority of our sample.

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

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

C.4 Morphing

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

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

D Appendix D: Randomized Lab Experiment

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

The causal interpretation of our new hypotheses is that heavy-faced or well-groomed defendants are released more often because these facial characteristics directly affect how

judges form judgments (consciously or unconsciously). Potential confounding arises from the fact that the judge has information that our algorithm does not (as we describe in Section 3), mainly what happens in the hearing itself. Mobius and Rosenblat (2006) show that people’s appearance can shape how confident they act, as well as their oral communication skills. Carrying that logic over to our application, it is possible that people who are more heavy-faced or well-groomed either act more confident in court (as signaled by for example their body language, eye contact with the judge or prosecutor, etc.), or are better able to explain themselves to either the judge or (more likely, since most defendants say little in court at pre-trial detention hearings) their own defense lawyer. These alternate mechanisms are interesting because they suggest different psychologies (and even implicate the psychologies of different people, e.g., the prosecutor or public defender rather than the judge).

We carry out a laboratory experiment that shuts down these two potential channels of confounding to isolate the independent causal effect of defendant appearance on judicial assessments of each defendant’s pre-trial risk. At a very high level we:

- 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 defendant 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 XXX study subjects on the M-turk platform. Our dataset is structured at the level of the respondent-and-defendant-pair. Our estimating equation is given as follows, with δ_{is} a set of defendant-pair fixed effects:

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

For our statistical analysis, we cluster the standard errors by respondent and defendant-pair as in Cameron et al. (2011). Conditioning on participant fixed effects yields very similar results.

We find ... (insert results here, start by showing that subjects use the Xs the way we expect, and then move on to discussing how they react to face info?)

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.⁸⁶ 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.

⁸⁶Recall the discussion in Section 4.2 argues against the possibility that these facial characteristics are proxies for risk.

Appendix Tables

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

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

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

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

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

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

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

Table A.III: Human Intelligence Tasks

Common Name	Survey Number	Short Description	Final Dataset	Subjects	Compensation	Additional Notes
Qualifying task	1	Subjects label 25 images across known variables, in order to identify high-quality raters.	Ratings from several MTurk workers, used to identify a qualified subpopulation of 343 MTurk workers.	600 MTurk Workers	8¢ per image	This survey was periodically re-run when a larger or updated population of raters was required. In total, 343 qualified MTurk Workers were identified across all qualification surveys.
Data collection labelling task part A	2	Subjects label 25 images on sliders for attractiveness, dominance, competence, trustworthiness and well-groomed, a free text input for age, a swatch for skin tone, and a selection for race.	Labels for 32881 images. Includes at least one label for age, race, and skin tone for all images in training and validation, at least three labels for all sliders in training dataset, and at least five labels for all sliders in validation dataset.	343 Qualified MTurk Workers	8¢ per image	The results from all labelling surveys was combined to produce a single image-label dataset.
Data collection labelling task part B	3	Subjects label 25 images on sliders for attractiveness, dominance, competence, trustworthiness, well-groomed, heavy-faced, and potentially other features.	See above.	343 Qualified MTurk Workers	5¢ per image	The results from this survey were merged with the other image label datasets.
Afro-centric features	4	Format is similar to survey 3, but sliders shown are for afrocentric features.	See above.	35 MTurk Workers from qualified population	5¢ per image	Workers were informed that the HIT contained "sensitive material". The results from this survey were merged with the other image label datasets.
Labelling quality check	5	Format is identical to survey 2, but each hit is repeated with multiple subjects.	100 images each with 10 labels.	40 MTurk Workers from qualified population	5¢ per image	The results from this survey were merged with the other image label datasets.
Human guess labelling task	6	Subjects are presented with 50 pairs of mugshots, and instructed to select which person they believe was detained. They are given feedback after each selection (so they can learn to identify patterns), and paid a 5 cent incentive for every correct guess. Each pair is matched to contain the same age bin, race, and sex.	Human guesses for 29,750 image pairs. The final dataset has at least three guesses for 8,001 images, with average of 7.4 guesses per image. Coverage is 79% for images.	595 Prolific Workers	\$3.00 base rate, plus a bonus of 5¢ for every correct guess	Because image pairs are matched on age bins, race, and sex, about 21 percent of our validation images do not have a proper match, and hence do not receive a human guess feature.
Morph labelling (along detention gradients)	7	Format and incentive is identical to survey 6, but image pairs shown are all morphed pairs with a high/low detain probability.	Comments described interpreted difference in image pairs, as seen by each Prolific worker. Also, guesses from each prolific worker to get a global masurement of accuracy.	54 Prolific Workers	\$3.00 base rate, plus a bonus of 5¢ for every correct guess	
Morph labelling (along residual gradients)	8	Format and incentive is identical to survey 6, but image pairs shown are all morph pairs with a high/low detain probability, and a similar estimated skin tone and well-groomed score.	Comments described interpreted difference in image pairs, as seen by each Prolific worker. Also, guesses from each prolific worker to get a global measurement of accuracy.	52 Prolific Workers	\$3.00 base rate, plus a bonus of 5¢ for every correct guess	
Morph labelling (along age gradients)	9	Format and incentive is identical to survey 6, but image pairs shown are all morph pairs with a high/low estimated age. Participants are not told what the "hidden characteristic" is, and must identify it from feedback.	Comments described interpreted difference in image pairs, as seen by each Prolific worker. Also, guesses from each prolific worker to get a global masurement of accuracy.	52 Prolific Workers	\$3.00 base rate, plus a bonus of 5¢ for every correct guess	

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

Population	<i>Mean Label Value</i>				
	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

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

	<i>Dependent variable:</i>				
	Algo Judge Detain Prediction		Judge Detain Decision		
	(1)	(2)	(3)	(4)	(5)
Heavy-Faced		-.0180*** (.0008)		-.0220*** (.0037)	-.0118*** (.0037)
Well-Groomed		-.0134*** (.0011)		-.0109** (.0051)	-.0033 (.0051)
Algo Judge Detain Prediction			.6065*** (.0443)		.5699*** (.0458)
Male	.1133*** (.0025)	.1120*** (.0024)	.0246** (.0118)	.0912*** (.0108)	.0274** (.0119)
Age	.0003*** (.0001)	.0004*** (.0001)	-.0014*** (.0004)	-.0011*** (.0004)	-.0013*** (.0004)
Black	-.0223*** (.0040)	-.0243*** (.0039)	-.0557*** (.0174)	-.0716*** (.0175)	-.0578*** (.0174)
Asian	-.0238** (.0112)	-.0166 (.0109)	-.0639 (.0487)	-.0714 (.0490)	-.0620 (.0487)
Indigenous American	.0107 (.0234)	.0011 (.0226)	.0645 (.1014)	.0578 (.1022)	.0571 (.1014)
Human Guess	.0387*** (.0061)	.0275*** (.0059)	.0840*** (.0266)	.0959*** (.0268)	.0803*** (.0267)
Constant	.0958*** (.0076)	.2731*** (.0108)	.0118 (.0333)	.2536*** (.0487)	.0980** (.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)	
Heavy-Faced		−.0193*** (.0010)		−.0106*** (.0014)		−.0191*** (.0043)	−.0078* (.0044)		−.0277*** (.0066)	−.0232*** (.0067)
Well-Groomed		−.0128*** (.0013)		−.0174*** (.0020)		−.0072 (.0060)	.0002 (.0059)		−.0254*** (.0097)	−.0179* (.0098)
Algo Judge Detain Prediction					.6027*** (.0505)		.5814*** (.0523)	.5376*** (.1020)		.4297*** (.1054)
Age	.0004*** (.0001)	.0006*** (.0001)	−.0003* (.0002)	−.0004** (.0002)	−.0014*** (.0005)	−.0010** (.0005)	−.0014*** (.0005)	−.0012 (.0008)	−.0016* (.0008)	−.0014* (.0008)
Black	−.0028 (.0048)	−.0068 (.0046)	−.0786*** (.0065)	−.0761*** (.0062)	−.0441** (.0209)	−.0494** (.0211)	−.0455** (.0209)	−.1018*** (.0309)	−.1394*** (.0298)	−.1067*** (.0308)
Asian	−.0091 (.0129)	−.0025 (.0124)	−.0625*** (.0209)	−.0544*** (.0202)	−.0536 (.0560)	−.0543 (.0565)	−.0528 (.0561)	−.0915 (.0963)	−.1106 (.0962)	−.0872 (.0960)
Indigenous American	.0173 (.0300)	.0087 (.0290)	−.0169 (.0316)	−.0251 (.0306)	−.0782 (.1307)	−.0780 (.1318)	−.0831 (.1307)	.3193** (.1456)	.2876** (.1458)	.2984** (.1452)
Human Guess	.0348*** (.0069)	.0247*** (.0067)	.0438*** (.0120)	.0281** (.0117)	.0678** (.0303)	.0809*** (.0306)	.0665** (.0303)	.1573*** (.0556)	.1512*** (.0558)	.1391** (.0556)
Constant	.1849*** (.0089)	.3630*** (.0126)	.1516*** (.0133)	.3190*** (.0189)	.0484 (.0399)	.3024*** (.0572)	.0914 (.0598)	−.0064 (.0630)	.3863*** (.0902)	.2492*** (.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

	<i>Dependent variable:</i>									
	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*** (.0010)		−.0166*** (.0014)		−.0210*** (.0044)	−.0112** (.0045)		−.0214*** (.0067)	−.0126* (.0068)
Well-Groomed		−.0184*** (.0014)		−.0046** (.0019)		−.0111* (.0062)	−.0008 (.0063)		−.0114 (.0090)	−.0090 (.0090)
Algo Judge Detain Prediction					.5915*** (.0532)		.5602*** (.0553)	.5690*** (.0852)		.5270*** (.0874)
Male	.1442*** (.0031)	.1415*** (.0030)	.0592*** (.0040)	.0607*** (.0039)	.0435*** (.0154)	.1245*** (.0135)	.0453*** (.0155)	−.0086 (.0189)	.0276 (.0183)	−.0045 (.0190)
Age	.0005*** (.0001)	.0005*** (.0001)	−.0002 (.0002)	−.00003 (.0002)	−.0013*** (.0005)	−.0010** (.0005)	−.0013*** (.0005)	−.0015** (.0008)	−.0015* (.0008)	−.0015* (.0008)
Human Guess	.0328*** (.0072)	.0224*** (.0070)	.0467*** (.0111)	.0349*** (.0109)	.0737** (.0312)	.0846*** (.0315)	.0720** (.0313)	.1037** (.0510)	.1124** (.0515)	.0940* (.0513)
Constant	.0514*** (.0172)	.2545*** (.0191)	.1445*** (.0121)	.2632*** (.0182)	.2020*** (.0745)	.4109*** (.0865)	.2683*** (.0870)	.0250 (.0567)	.2961*** (.0858)	.1574* (.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

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

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

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

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

	<i>Dependent variable:</i>									
	Judge Detain Decision						MI $\geq$ Median(MI)		MI < Median(MI)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Heavy-Faced	-0.0223*** (0.0036)	-0.0214*** (0.0037)			-0.0210*** (0.0037)		-0.0229*** (0.0051)		-0.0190*** (0.0054)	
Well-Groomed			-0.0168*** (0.0044)	-0.0106** (0.0052)	-0.0087* (0.0052)		-0.0135* (0.0072)		-0.0039 (0.0075)	
Algorithmic Judge Prediction						0.6109*** (0.0436)		0.4845*** (0.0602)		0.7695*** (0.0632)
Male		0.0939*** (0.0108)		0.0981*** (0.0108)	0.0946*** (0.0108)	0.0266** (0.0118)	0.0897*** (0.0147)	0.0352** (0.0161)	0.1010*** (0.0159)	0.0145 (0.0173)
Age		-0.0012*** (0.0004)		-0.0014*** (0.0004)	-0.0012*** (0.0004)	-0.0015*** (0.0004)	-0.0011* (0.0006)	-0.0014*** (0.0005)	-0.0014** (0.0006)	-0.0014** (0.0006)
Black		-0.0634*** (0.0156)		-0.0611*** (0.0156)	-0.0633*** (0.0156)	-0.0514*** (0.0154)	-0.0484** (0.0219)	-0.0409* (0.0218)	-0.0793*** (0.0222)	-0.0640*** (0.0218)
Asian		-0.0718 (0.0488)		-0.0707 (0.0489)	-0.0689 (0.0488)	-0.0623 (0.0484)	-0.0484 (0.0775)	-0.0385 (0.0772)	-0.0850 (0.0625)	-0.0807 (0.0615)
Indigenous American		0.0505 (0.1019)		0.0687 (0.1020)	0.0533 (0.1019)	0.0575 (0.1010)	0.0472 (0.1614)	0.0596 (0.1607)	0.0551 (0.1308)	0.0387 (0.1287)
Skin-Tone		-0.1047*** (0.0250)		-0.1019*** (0.0251)	-0.1061*** (0.0250)	-0.0754*** (0.0249)	-0.0810** (0.0352)	-0.0554 (0.0351)	-0.1354*** (0.0357)	-0.0994*** (0.0352)
Attractiveness		-0.0070 (0.0068)		0.0012 (0.0070)	-0.0037 (0.0070)	-0.0002 (0.0067)	-0.0048 (0.0100)	-0.0045 (0.0095)	-0.0029 (0.0099)	0.0043 (0.0093)
Competence		-0.0183** (0.0072)		-0.0165** (0.0073)	-0.0169** (0.0073)	-0.0136* (0.0072)	-0.0222** (0.0102)	-0.0201** (0.0101)	-0.0110 (0.0104)	-0.0072 (0.0102)
Dominance		0.0101* (0.0052)		0.0067 (0.0052)	0.0101* (0.0052)	0.0052 (0.0051)	0.0178** (0.0072)	0.0123* (0.0071)	0.0016 (0.0075)	-0.0032 (0.0074)
Trustworthiness		-0.0081 (0.0071)		-0.0099 (0.0071)	-0.0072 (0.0071)	-0.0088 (0.0070)	-0.0058 (0.0102)	-0.0086 (0.0101)	-0.0099 (0.0099)	-0.0103 (0.0097)
Human Guess		0.0986*** (0.0268)		0.1009*** (0.0268)	0.0958*** (0.0268)	0.0824*** (0.0266)	0.0991*** (0.0380)	0.0957** (0.0378)	0.0929** (0.0378)	0.0672* (0.0373)
Mental Illness (MI)	0.0103*** (0.0033)	0.0073** (0.0035)	0.0088** (0.0034)	0.0088** (0.0035)	0.0065* (0.0035)	0.0055 (0.0034)				
Constant	0.3099*** (0.0248)	0.4032*** (0.0486)	0.2783*** (0.0285)	0.3165*** (0.0469)	0.4276*** (0.0507)	0.1724*** (0.0445)	0.4530*** (0.0643)	0.2076*** (0.0593)	0.4560*** (0.0680)	0.1821*** (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.XIV: Relationship between novel features and judge decision controlling for defendant's perceived socio-economic status (SES)

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

Appendix Figures

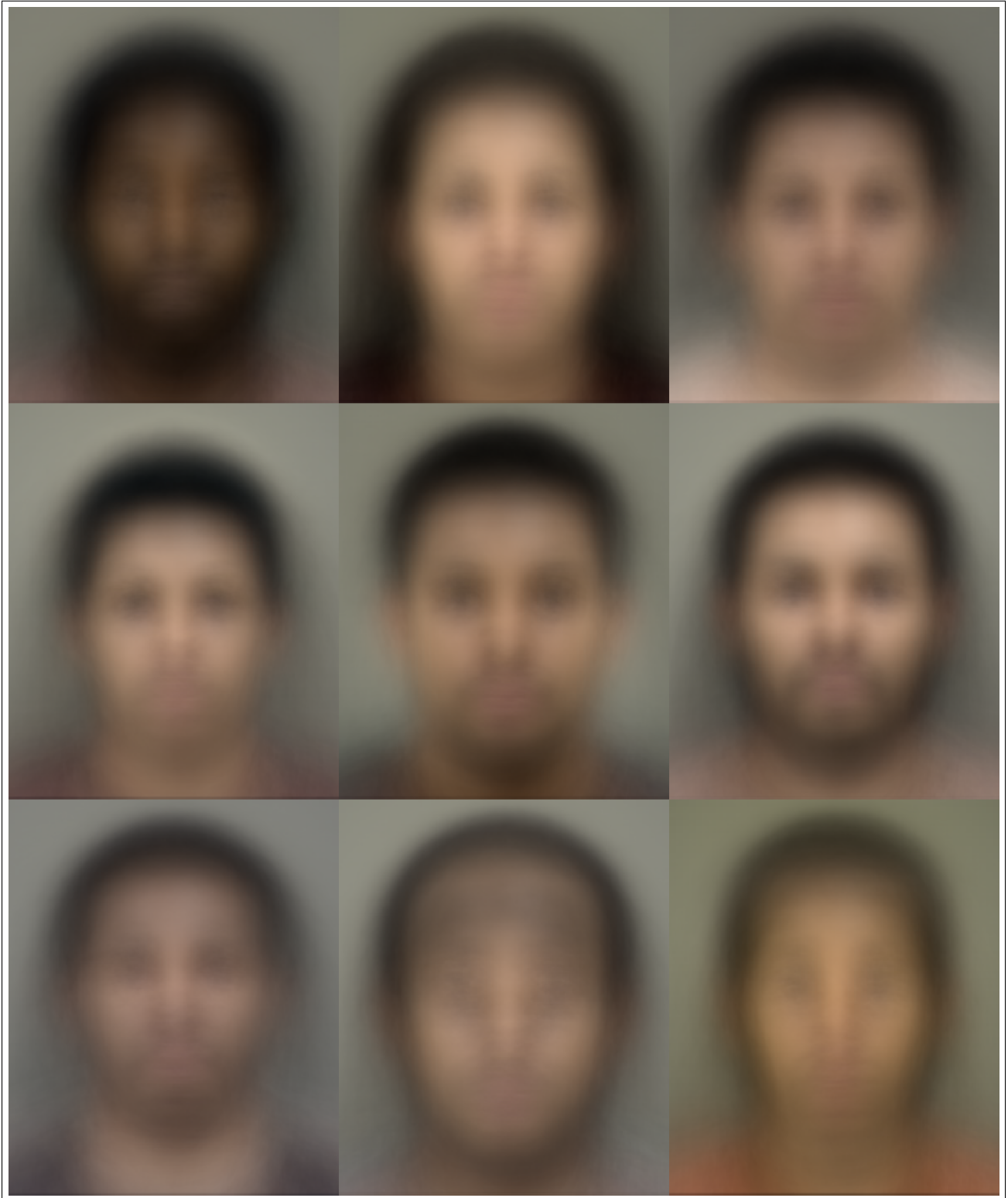


Figure A.I: Eigenfaces

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

Consent for Participation in Research study

Study Title: AI and Judges

Principal Investigator: Dr. Jens Ludwig

IRB Study Number: IRB20-0917

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

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

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

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

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

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

Contacts & Questions:

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

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

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

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

Consent:

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

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

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

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

Instructions

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

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

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

In addition:

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

Start Survey >

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

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



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 unappealing 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.III: 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.

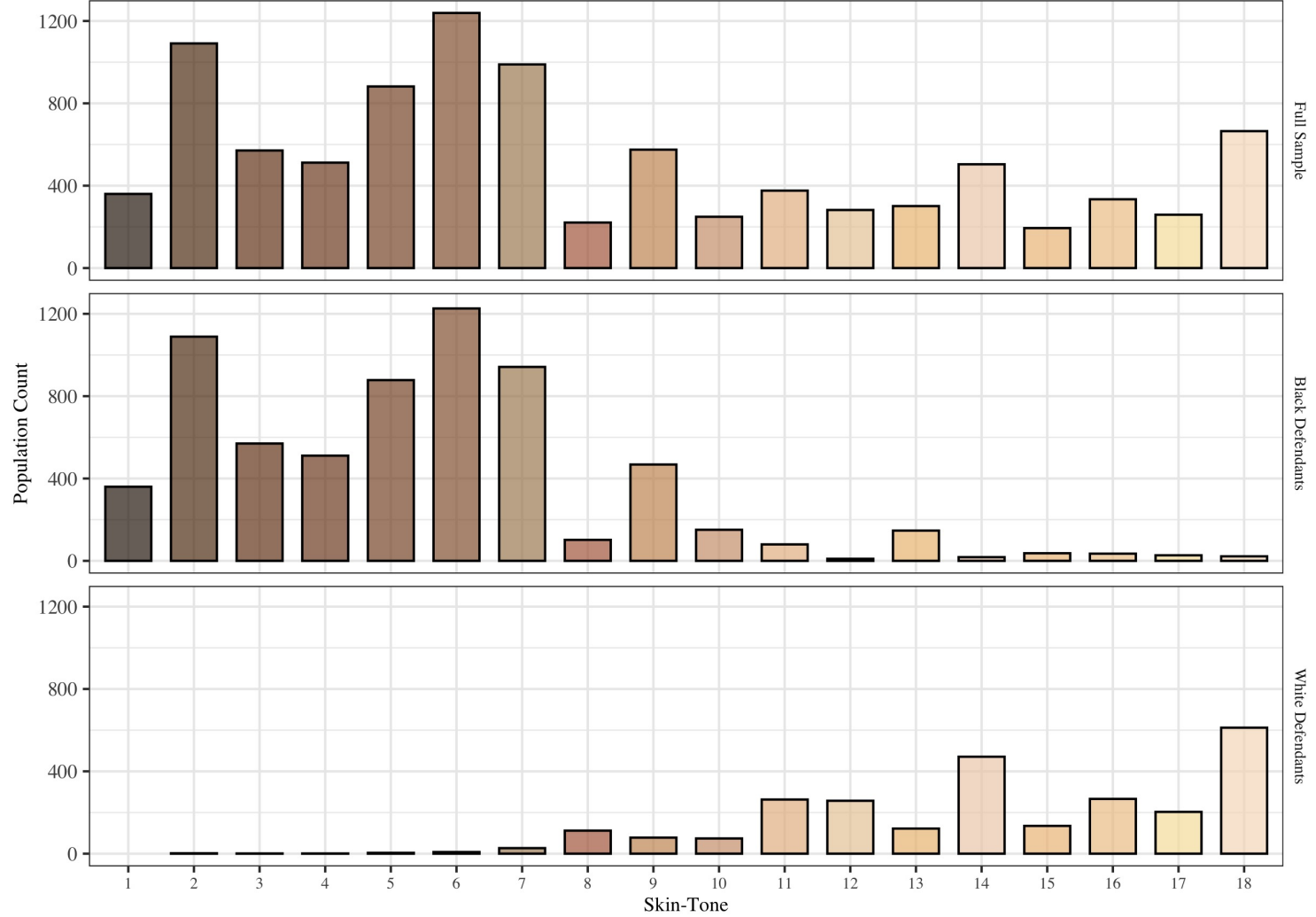


Figure A.IV: Distribution of skin-tone categories for full validation sample, and by defendant race

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.

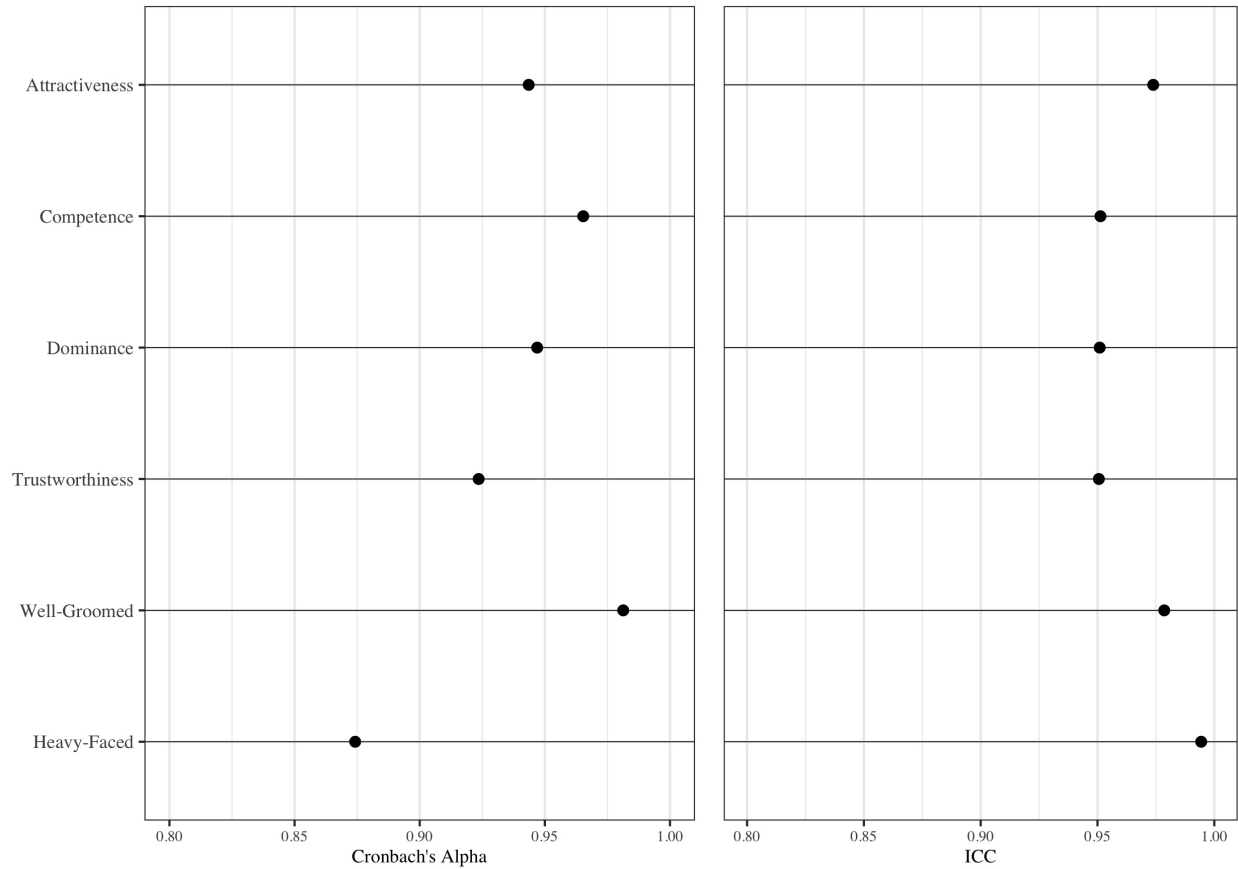


Figure A.V: Reliability measures for human-rated psychological features

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.

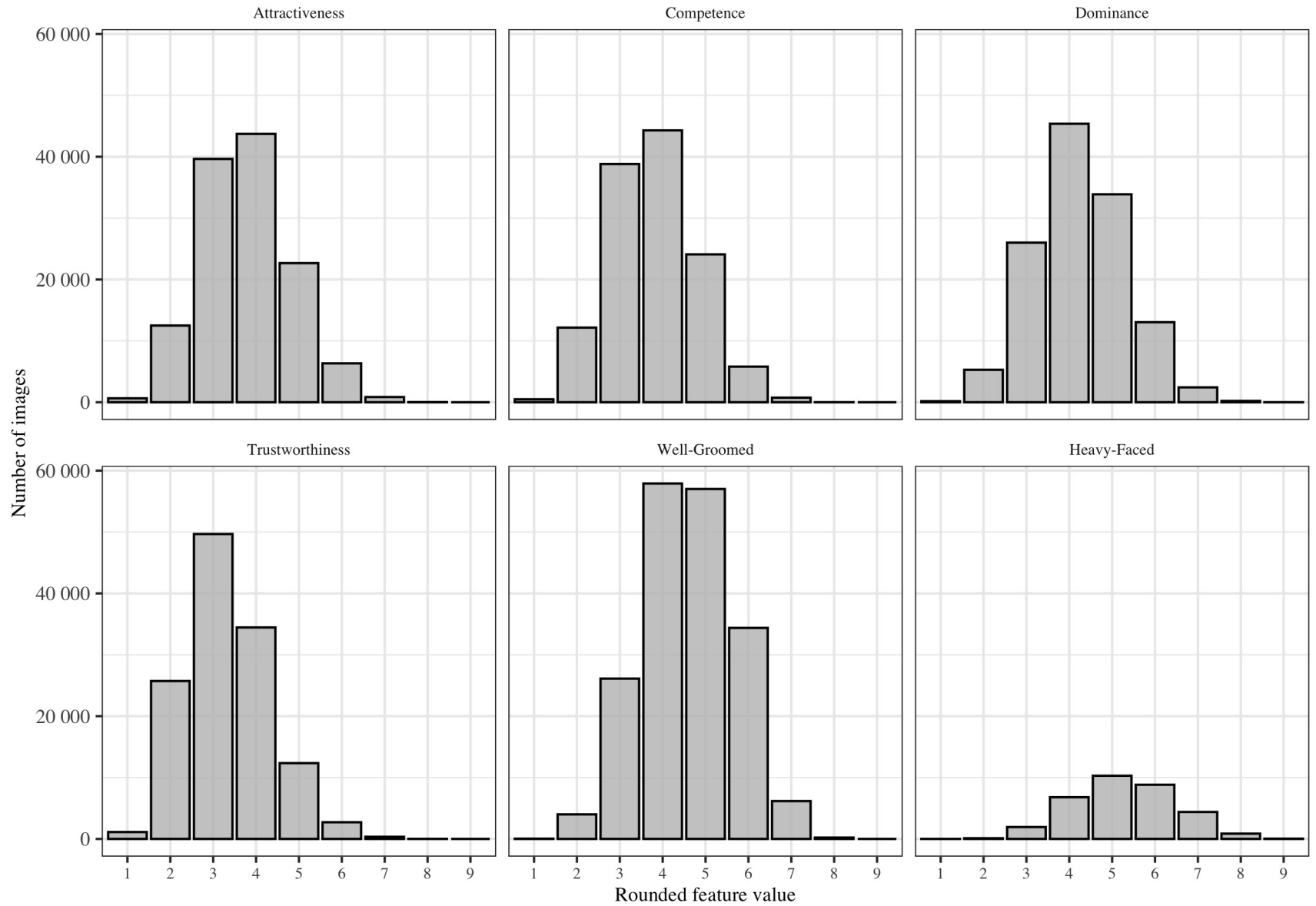


Figure A.VI: Distribution of human ratings of psychological features based on face images

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).

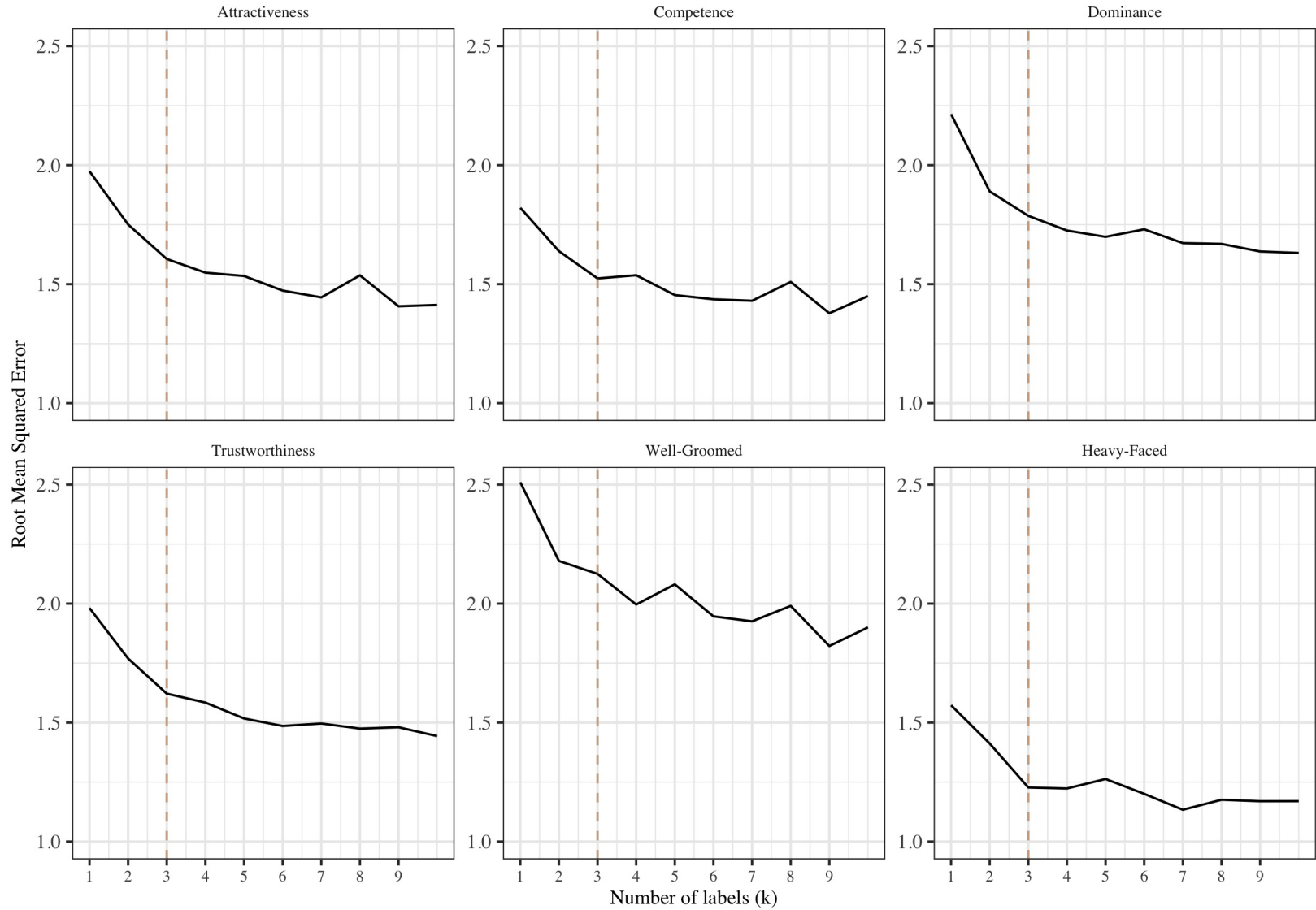


Figure A.VII: Signal vs. noise in human ratings by number of ratings provided

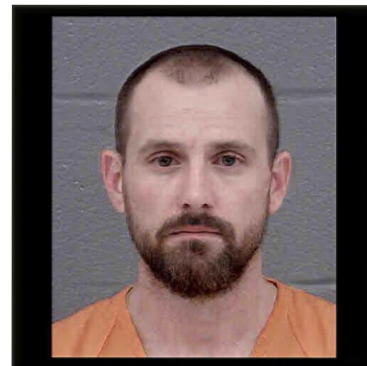
Notes: The figure shows the results of taking the average of the first K labels provided by human raters for that psychological feature from looking at a face image, and using that to predict the value of the next $(K + 1)$ human rating of that same image on the same psychological feature, reported in 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.**

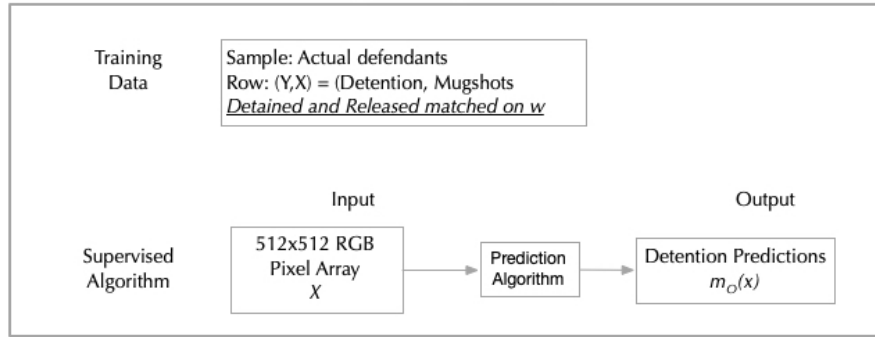


(b) The screen presented to workers after selecting an image. In addition to the green outline, a popup window appeared informing candidates if their selection was correct.

Figure A.VIII: 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.

Orthogonalized Judge Detention Predictor Orthogonalizing with respect to w



Orthogonalized Morphing Step (Orthogonalizing with respect to w)

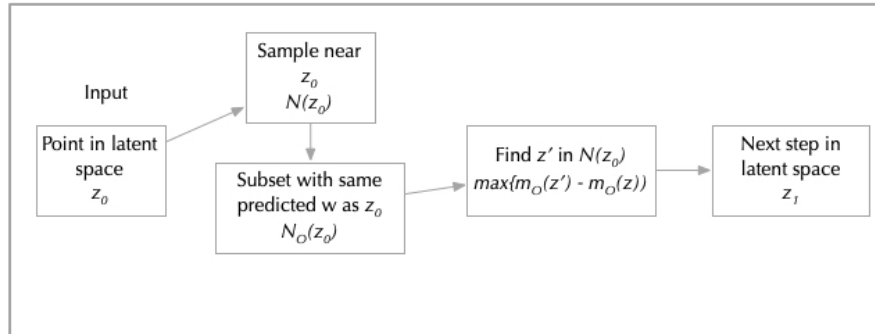


Figure A.IX: Orthogonalization pipeline

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 unattractive, 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.X: Example of instructions given to M-turkers for one of a labelling task

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.XI: 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:

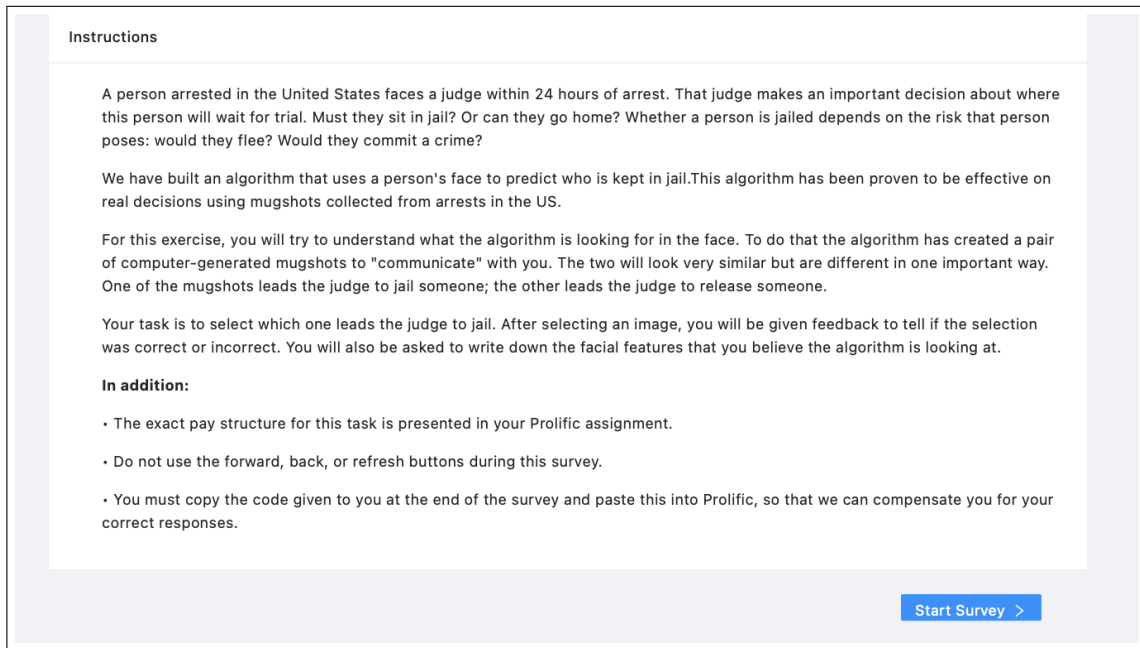
We want to include opposites as the same feature. For example, stressed face / relaxed face is the same feature, since they are opposites; long hair / short hair are the same feature, but not the same feature as curly hair / straight hair; neutral face / happy face are not really the same feature, since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

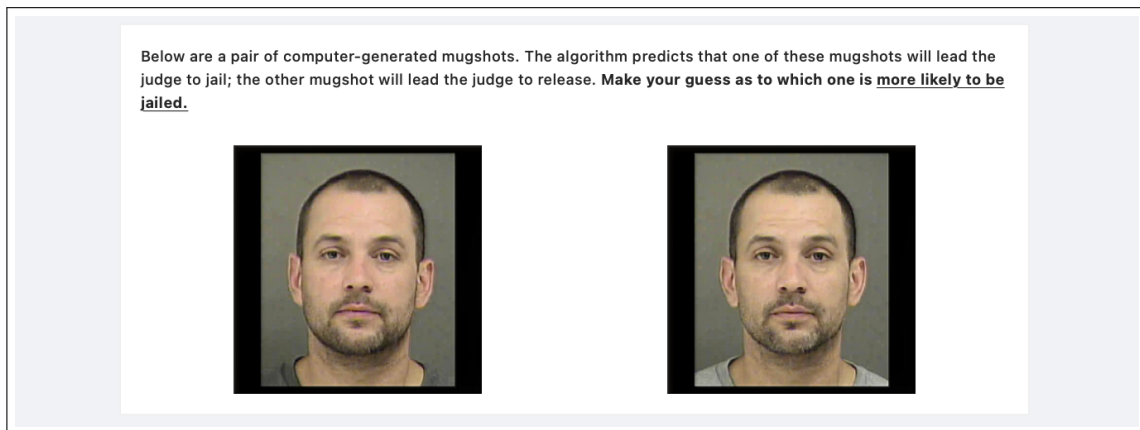
We are looking for features that are specific and measurable. A good rule of thumb is: good features are something about a face, bad features are just parts of a face.

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.XII: Instructions shown to independent RAs for the comment categorization task



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



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

Figure A.XIII: 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.

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.XIV: 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.

References

- Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Biral Runesha, and Teodora Szasz, “What we teach about race and gender: Representation in images and text of children’s books,” Technical Report, National Bureau of Economic Research 2021.
- Agan, Amanda Y, Jennifer L Doleac, and Anna Harvey, “Misdemeanor prosecution,” Technical Report, National Bureau of Economic Research 2021.
- Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin, “Learning certifiably optimal rule lists for categorical data,” *Journal of Machine Learning Research*, 2018, 18, 1–78.
- Angelova, Victoria, Will Dobbie, and Crystal S Yang, “Algorithmic Recommendations and Human Discretion,” 2022.
- Arnold, David, Will Dobbie, and Crystal S Yang, “Racial bias in bail decisions,” *The Quarterly Journal of Economics*, 2018, 133 (4), 1885–1932.
- , Will S Dobbie, and Peter Hull, “Measuring racial discrimination in bail decisions,” Technical Report, National Bureau of Economic Research 2020.
- Athey, Susan, “Beyond prediction: Using big data for policy problems,” *Science*, 2017, 355 (6324), 483–485.
- , “The impact of machine learning on economics,” in “The economics of artificial intelligence: An agenda,” University of Chicago Press, 2018, pp. 507–547.
- and Guido W Imbens, “Machine learning methods that economists should know about,” *Annual Review of Economics*, 2019, 11, 685–725.
- , Dean Karlan, Emil Palikot, and Yuan Yuan, “Smiles in profiles: Improving fairness and efficiency using estimates of user preferences in online marketplaces,” Technical Report, National Bureau of Economic Research 2022.
- , Guido W Imbens, Jonas Metzger, and Evan Munro, “Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations,” *Journal of Econometrics*, 2021.
- Autor, David, “Polanyi’s paradox and the shape of employment growth,” Technical Report, National Bureau of Economic Research 2014.
- Avitzour, Eliana, Adi Choen, Daphna Joel, and Victor Lavy, “On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes,” Technical Report, National Bureau of Economic Research 2020.
- Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller, “How to explain individual classification decisions,” *The Journal of Machine Learning Research*, 2010, 11 (1), 1803–1831.
- Bai, Xiao, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim, “Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments,” *Pattern Recognition*, 2021, 120, 108102.
- Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, 2019, 41 (2), 423–443.
- Begall, Sabine, Jaroslav Červený, 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, Naf-tali 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 birth-days,” *Available at SSRN 3203624*, 2020.
- , **Tobias J Moskowitz, and Kelly Shue**, “Decision making under the gambler’s fallacy: Evidence from asylum judges, loan officers, and baseball umpires,” *The Quarterly Journal of Economics*, 2016, *131* (3), 1181–1242.
- Dahl, Gordon B and Matthew M Knepper**, “Age discrimination across the business cycle,” Technical Report, National Bureau of Economic Research 2020.
- Davies, Alex, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász et al.**, “Advancing mathematics by guiding human intuition with AI,” *Nature*, 2021, *600* (7887), 70–74.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- Dobbie, Will and Crystal S Yang**, “The US pretrial system: Balancing individual rights and

- public interests,” *Journal of Economic Perspectives*, 2021, 35 (4), 49–70.
- , **Jacob Goldin**, and **Crystal S. Yang**, “The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges,” *American Economic Review*, February 2018, 108 (2), 201–240.
- Doshi-Velez, Finale** and **Been Kim**, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- Eberhardt, Jennifer L**, **Paul G Davies**, **Valerie J Purdie-Vaughns**, and **Sheri Lynn Johnson**, “Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes,” *Psychological science*, 2006, 17 (5), 383–386.
- Einav, Liran** and **Jonathan Levin**, “The data revolution and economic analysis,” *Innovation Policy and the Economy*, 2014, 14 (1), 1–24.
- Eren, Ozkan** and **Naci Mocan**, “Emotional judges and unlucky juveniles,” *American Economic Journal: Applied Economics*, 2018, 10 (3), 171–205.
- Freitas, Alex A**, “Comprehensible classification models: a position paper,” *ACM SIGKDD explorations newsletter*, 2014, 15 (1), 1–10.
- Freund, Yoav**, **Robert Schapire**, and **Naoki Abe**, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, 1999, 14 (5), 771–780.
- Frieze, Irene Hanson**, **Josephine E Olson**, and **June Russell**, “Attractiveness and income for men and women in management,” *Journal of Applied Social Psychology*, 1991, 21 (13), 1039–1057.
- Fudenberg, Drew** and **Annie Liang**, “Predicting and understanding initial play,” *American Economic Review*, 2019, 109 (12), 4112–4141.
- Gentzkow, Matthew**, **Bryan Kelly**, and **Matt Taddy**, “Text as data,” *Journal of Economic Literature*, 2019, 57 (3), 535–74.
- Ghandeharioun, Asma**, **Been Kim**, **Chun-Liang Li**, **Brendan Jou**, **Brian Eoff**, and **Rosalind W Picard**, “Dissect: Disentangled simultaneous explanations via concept traversals,” *arXiv preprint arXiv:2105.15164*, 2021.
- Ghorbani, Amirata**, **James Wexler**, **James Y Zou**, and **Been Kim**, “Towards automatic concept-based explanations,” *Advances in Neural Information Processing Systems*, 2019, 32.
- Goldin, Claudia** and **Cecilia Rouse**, “Orchestrating impartiality: The impact of" blind" auditions on female musicians,” *American economic review*, 2000, 90 (4), 715–741.
- Goncalves, Felipe** and **Steven Mello**, “A few bad apples? Racial bias in policing,” *American Economic Review*, 2021, 111 (5), 1406–1441.
- Goodfellow, Ian J**, **Jonathon Shlens**, and **Christian Szegedy**, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- Goodfellow, Ian**, **Jean Pouget-Abadie**, **Mehdi Mirza**, **Bing Xu**, **David Warde-Farley**, **Sherjil Ozair**, **Aaron Courville**, and **Yoshua Bengio**, “Generative adversarial nets,” *Advances in neural information processing systems*, 2014, 27, 2672–2680.
- , –, –, –, –, –, –, –, and –, “Generative adversarial networks,” *Communications of the ACM*, 2020, 63 (11), 139–144.
- 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 Adalı, and Christian Jutten**, “Multimodal data fusion: an overview of methods, challenges, and prospects,” *Proceedings of the IEEE*, 2015, 103 (9), 1449–1477.
- Lang, Oran, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani et al.**, “Explaining in style: Training a gan to explain a classifier in stylespace,” in “Proceedings of the IEEE/CVF International Conference on Computer Vision” 2021, pp. 693–702.
- Langley, Pat, Herbert A Simon, Gary L Bradshaw, and Jan M Zytkow**, *Scientific discovery*, Cambridge, Ma: MIT Press, 1987.
- LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet**, “Convolutional networks and applications in vision,” in “Proceedings of 2010 IEEE international symposium on circuits and systems” IEEE 2010, pp. 253–256.
- , **Yoshua Bengio, and Geoffrey Hinton**, “Deep learning,” *nature*, 2015, 521 (7553), 436–444.
- Lee, Minhyeok and Junhee Seok**, “Controllable generative adversarial network,” *Ieee Access*, 2019, 7, 28158–28169.
- Leskovec, Jure, Lars Backstrom, and Jon Kleinberg**, “Meme-tracking and the dynamics of the news cycle,” in “Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining” 2009, pp. 497–506.
- Letham, Benjamin, Cynthia Rudin, Tyler H McCormick, and David Madigan**, “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model,” *The Annals of Applied Statistics*, 2015, 9 (3), 1350–1371.
- Li, Oscar, Hao Liu, Chaofan Chen, and Cynthia Rudin**, “Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions,” in “Proceedings of the AAAI Conference on Artificial Intelligence,” Vol. 32 2018, pp. 3530–3537.
- Little, Anthony C, Benedict C Jones, and Lisa M DeBruine**, “Facial attractiveness: evolutionary based research,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2011, 366 (1571), 1638–1659.
- Liu, Shusen, Bhavya Kailkhura, Donald Loveland, and Yong Han**, “Generative counterfactual introspection for explainable deep learning,” in “2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)” IEEE 2019, pp. 1–5.
- Marcinkevičs, Ričards and Julia E Vogt**, “Interpretability and explainability: A machine learning zoo mini-tour,” *arXiv preprint arXiv:2012.01805*, 2020.
- Miller, Andrew, Ziad Obermeyer, John Cunningham, and Sendhil Mullainathan**, “Discriminative regularization for latent variable models with applications to electrocardiography,” in “International Conference on Machine Learning” PMLR 2019, pp. 8072–8081.
- Mobius, Markus M and Tanya S Rosenblat**, “Why beauty matters,” *American Economic Review*, 2006, 96 (1), 222–235.
- Mobley, R Keith**, *An introduction to predictive maintenance*, Elsevier, 2002.
- Mullainathan, Sendhil and Jann Spiess**, “Machine learning: an applied econometric approach,” *Journal of Economic Perspectives*, 2017, 31 (2), 87–106.
- and **Ziad Obermeyer**, “Diagnosing physician error: A machine learning approach to low-value health care,” *The Quarterly Journal of Economics*, 2022, 137 (2), 679–727.

- Murphy, Allan H**, “A new vector partition of the probability score,” *Journal of Applied Meteorology and Climatology*, 1973, 12 (4), 595–600.
- Nalisnick, Eric, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan**, “Do deep generative models know what they don’t know?,” *arXiv preprint arXiv:1810.09136*, 2018.
- Narayanaswamy, Arunachalam, Subhashini Venugopalan, Dale R Webster, Lily Peng, Greg S Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C Nelson, and Avinash V Varadarajan**, “Scientific discovery by generating counterfactuals using image translation,” in “International Conference on Medical Image Computing and Computer-Assisted Intervention” Springer 2020, pp. 273–283.
- Neumark, David, Ian Burn, and Patrick Button**, “Experimental age discrimination evidence and the Heckman critique,” *American Economic Review*, 2016, 106 (5), 303–308.
- Nielsen, Michael A**, *Neural networks and deep learning*, Vol. 25, Determination press San Francisco, CA, 2015.
- Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune**, “Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning,” *Proceedings of the National Academy of Sciences*, 2018, 115 (25), E5716–E5725.
- Oosterhof, Nikolaas N and Alexander Todorov**, “The functional basis of face evaluation,” *Proceedings of the National Academy of Sciences*, 2008, 105 (32), 11087–11092.
- Peterson, Joshua C, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths**, “Using large-scale experiments and machine learning to discover theories of human decision-making,” *Science*, 2021, 372 (6547), 1209–1214.
- , **Stefan Uddenberg, Thomas L Griffiths, Alexander Todorov, and Jordan W Suchow**, “Deep models of superficial face judgments,” *Proceedings of the National Academy of Sciences*, 2022, 119 (17), e2115228119.
- Pierson, Emma, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer**, “An algorithmic approach to reducing unexplained pain disparities in underserved populations,” *Nature Medicine*, 2021, 27 (1), 136–140.
- Pion-Tonachini, Luca, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, Benjamin Nachman et al.**, “Learning from learning machines: a new generation of AI technology to meet the needs of science,” *arXiv preprint arXiv:2111.13786*, 2021.
- Popper, Karl**, *The logic of scientific discovery*, Routledge, 2005.
- Pronin, Emily**, “The introspection illusion,” *Advances in experimental social psychology*, 2009, 41, 1–67.
- Ramachandram, Dhanesh and Graham W Taylor**, “Deep multimodal learning: A survey on recent advances and trends,” *IEEE signal processing magazine*, 2017, 34 (6), 96–108.
- Rambachan, Ashesh et al.**, “Identifying prediction mistakes in observational data,” *Harvard University*, 2021.
- Rawat, Waseem and Zenghui Wang**, “Deep convolutional neural networks for image classification: A comprehensive review,” *Neural computation*, 2017, 29 (9), 2352–2449.
- Redcross, Cindy, Britt Henderson, L Miratrix, and E Valentine**, “Evaluation of pretrial justice system reforms that use the Public Safety Assessment: Effects in Mecklenburg County North Carolina Report 2,” *MDRC Center for Criminal Justice Research*. https://www.mdrc.org/sites/default/files/PSA_Mecklenburg_Brief2.pdf, 2019.

- Rudin, Cynthia**, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, 2019, 1 (5), 206–215.
- , **Rebecca J Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac**, “A process for predicting manhole events in Manhattan,” *Machine Learning*, 2010, 80 (1), 1–31.
- Said-Metwaly, Sameh, Wim Van den Noortgate, and Eva Kyndt**, “Approaches to measuring creativity: A systematic literature review,” *Creativity. Theories–Research–Applications*, 2017, 4 (2), 238–275.
- Sajjadi, Mehdi SM, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly**, “Assessing generative models via precision and recall,” *Advances in neural information processing systems*, 2018, 31.
- Schickore, Jutta**, “Scientific Discovery,” in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*, summer 2018 ed., Metaphysics Research Lab, Stanford University, 2018.
- Schlag, Pierre**, “Law and phrenology,” *Harvard Law Review*, 1997, 110 (4), 877–921.
- Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman**, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in “In Workshop at International Conference on Learning Representations” Citeseer 2014.
- Sirovich, Lawrence and Michael Kirby**, “Low-dimensional procedure for the characterization of human faces,” *Josa a*, 1987, 4 (3), 519–524.
- Sunstein, Cass R**, “Governing by algorithm? No noise and (potentially) less bias,” *Duke LJ*, 2021, 71, 1175.
- Swanson, Don R**, “Fish oil, Raynaud’s syndrome, and undiscovered public knowledge,” *Perspectives in biology and medicine*, 1986, 30 (1), 7–18.
- , “Migraine and magnesium: eleven neglected connections,” *Perspectives in biology and medicine*, 1988, 31 (4), 526–557.
- Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus**, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- Todorov, Alexander and DongWon Oh**, “The structure and perceptual basis of social judgments from faces,” in “Advances in experimental social psychology,” Vol. 63, Elsevier, 2021, pp. 189–245.
- , **Christopher Y Olivola, Ron Dotsch, Peter Mende-Siedlecki et al.**, “Social attributions from faces: Determinants, consequences, accuracy, and functional significance,” *Annual review of psychology*, 2015, 66 (1), 519–545.
- Ustun, Berk and Cynthia Rudin**, “Learning Optimized Risk Scores,” *Journal of Machine Learning Research*, 2019, 20 (150), 1–75.
- Varian, Hal R**, “Big data: New tricks for econometrics,” *Journal of Economic Perspectives*, 2014, 28 (2), 3–28.
- Wachter, Sandra, Brent Mittelstadt, and Chris Russell**, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” *Harvard Journal of Law & Technology*, 2018, 31 (2), 841–888.
- Wilson, Timothy D**, *Strangers to ourselves*, Harvard University Press, 2004.
- Yegnanarayana, Bayya**, *Artificial neural networks*, PHI Learning Pvt. Ltd., 2009.
- Yuhas, Ben P, Moise H Goldstein, and Terrence J Sejnowski**, “Integration of acoustic and visual speech signals using neural networks,” *IEEE Communications Magazine*, 1989, 27 (11),

65–71.

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.