MACHINE LEARNING AS A TOOL FOR HYPOTHESIS GENERATION*

JENS LUDWIG AND SENDHIL MULLAINATHAN

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

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

© The Author(s) 2024. Published by Oxford University Press on behalf of the President and Fellows of Harvard College. All rights reserved. For Permissions, please email: journals.permissions@oup.com
The Quarterly Journal of Economics (2024), 1–77. https://doi.org/10.1093/qje/qjad055.
Advance Access publication on January 10, 2024.

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

I. Introduction

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

FIGURE I

Correlates of Judge Detention Decision and Algorithmic Prediction of Judge Decision

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

FIGURE I

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

respect to detention odds, to minimize extraneous differences and help focus on what truly matters for judge detention decisions.

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

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

What are the two discovered features? The first can be called "well-groomed" (e.g., tidy, clean, groomed, versus unkept, disheveled, sloppy look), and the second can be called "heavy-faced" (e.g., wide facial shape, puffier face, wider face, rounder

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

First, to generate interpretable hypotheses, we must overcome the notorious black box nature of most machine learning algorithms. Unlike with a regression, one cannot simply inspect the coefficients. A modern deep-learning algorithm, for example, can have tens of millions of parameters. Noninspectability is

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

II. A Simple Framework for Discovery

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

II.A. The Goals of Hypothesis Generation

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

II.B. Human versus Algorithmic Hypothesis Generation

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

II.C. Related Methods

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

III. Application and Data

III.A. Judicial Decision Making

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

III.B. Administrative Data

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

FIGURE II

Illustrative Facial Images

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE I

SUMMARY STATISTICS FOR MECKLENBURG COUNTY NC DATA, 2017–2020

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE I
CONTINUED

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

Notes. This table reports descriptive statistics for our full data set and analysis subsets, which cover the period January 18, 2017, through January 17, 2020, from Mecklenburg County, NC. The lock-box hold-out data set consists of data from the last six months of our study period (July 17, 2019–January 17, 2020) plus a subset of cases through July 16, 2019, selected by randomly selecting arrestees. The remainder of the data set is then randomly assigned by arrestee to our training data set (used to build our algorithms) or to our validation set (which we use to report results in the article's main exhibits). For additional details of our data filters and partitioning procedures, see Online Appendix Table A.I. We define pretrial release as being released on the defendant's own recognizance or having been assigned and then posting cash bail requirements within three days of arrest. We define rearrest as experiencing a new arrest before adjudication of the focal arrest, with detained defendants being assigned zero values for the purposes of this table. Arrest charge categories reflect the most serious criminal charge for which a person was arrested, using the FBI Uniform Crime Reporting hierarchy rule in cases where someone is arrested and charged with multiple offenses. For analyses of variance for the test of the joint null hypothesis that the difference in means across each variable is zero, see Online Appendix Table A.II.

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

III.C. Human Labels

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

IV. The Surprising Importance of the Face

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

IV.A. What Drives Judge Decisions?

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

IV.B. Judicial Error?

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

seem to be strongly related enough to rearrest risk to explain the judge's use of it in making detention decisions.[42]

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

IV.C. Is the Algorithm Discovering Something New?

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE II

IS THE ALGORITHM REDISCOVERING KNOWN FACIAL FEATURES?

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE III

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE III
CONTINUED

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

V. Algorithm-Human Communication

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

V.A. The Challenge of Explanation

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

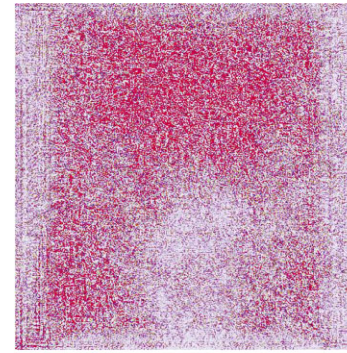
Even if saliency maps were more selective in highlighting pixels in applications like ours, for hypothesis generation they also suffer from a second limitation: they do not convey enough information to enable people to articulate interpretable hypotheses. In the cat detector example, a saliency map can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

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

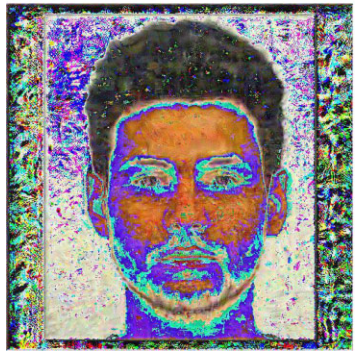
51. As noted already, to avoid contributing to the stereotyping of minorities in discussions of crime, in our exhibits we show images for non-Hispanic white men, although in our HITs we use images representative of the larger defendant population.

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

(A) Initial face



(B) Saliency map



(C) Naive age-morphed image

(D) Morphs from our procedure

FIGURE III

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

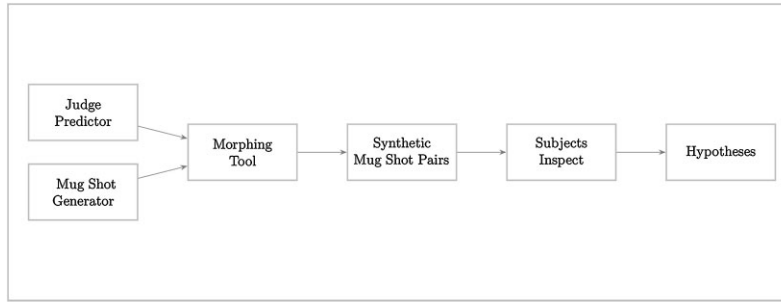


FIGURE IV

Hypothesis Generation Pipeline

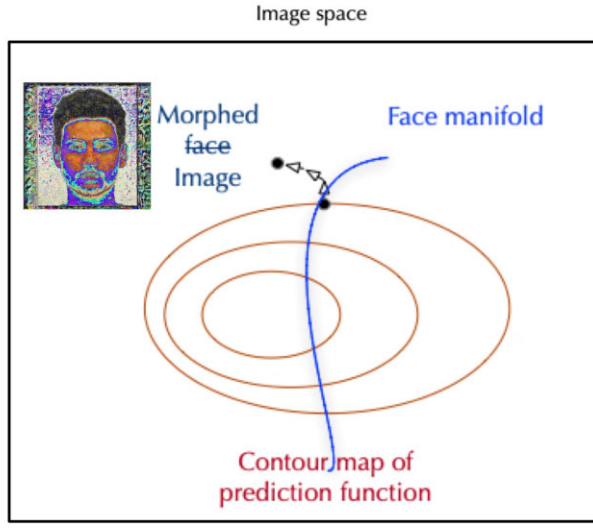
The diagram illustrates all the algorithmic components in our procedure by presenting a full pipeline for algorithmic interpretation.

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

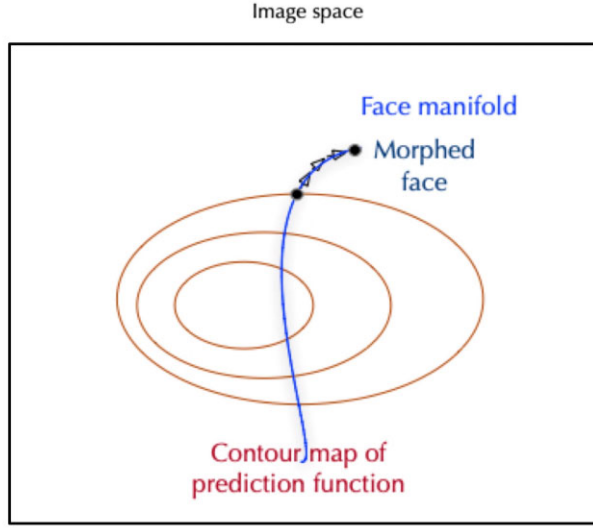
Figure V illustrates how this morphing procedure works in practice and highlights some of the technical challenges that arise. Let the box in the top panel represent the space of all possible images—all possible combinations of pixel values for, say, a 512×512 image. Within this space, we can apply our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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024



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



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

FIGURE V

Morphing Images for Detention Risk On and Off the Face Manifold

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

FIGURE V

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

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

V.B. Building a Model of the Data Distribution

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

are used to train the discriminator, which results in the generator being trained as it seeks to fool the discriminator. With machine learning, the performance of C and G improve with successive iterations of training. A perfect G would output images where the classifier C does no better than random guessing. Such a generator would by definition limit itself to the same input space that defines real images, that is, the data distribution of faces. (Additional discussion of GANs in general and how we construct our GAN specifically are in Online Appendix B.)

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

both detention rates and the type of background they have someone stand in front of for mug shots).

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

V.C. Validating the Morphing Procedure

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

V.D. Understanding the Judge Detention Predictor

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

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

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

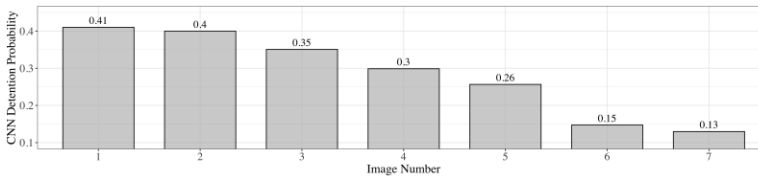
56. In principle this semantic grouping could be carried out in other ways, for example, with automated procedures involving natural-language processing.

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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



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

FIGURE VI

Illustration of Morphed Faces along the Detention Gradient

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

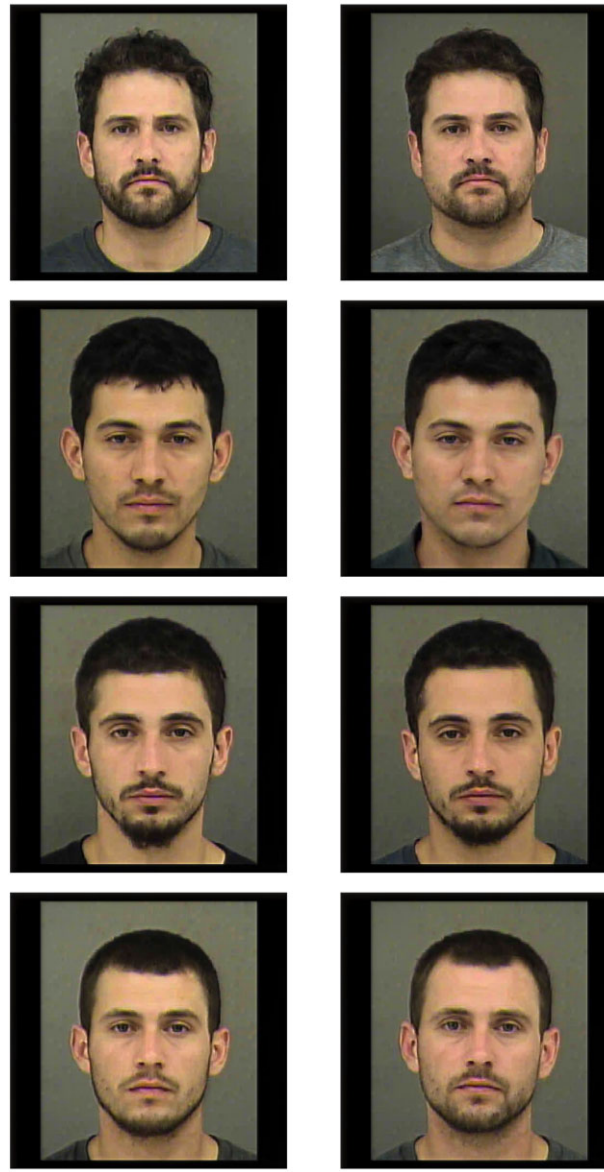
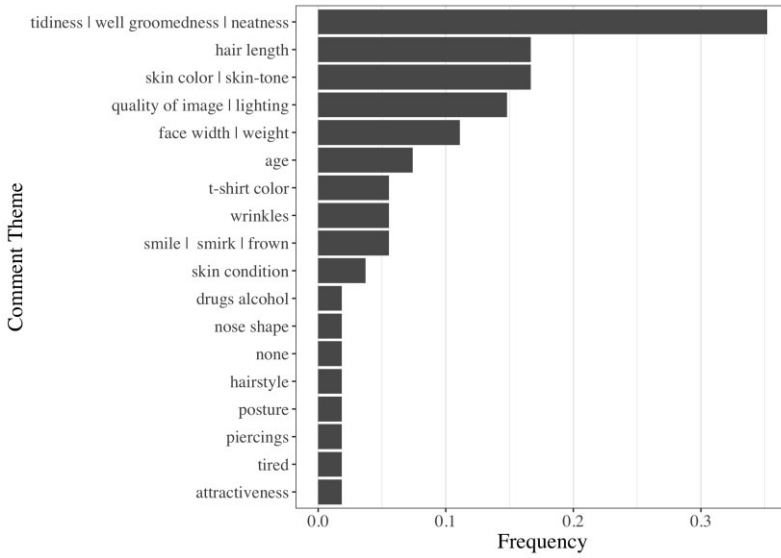


FIGURE VII

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

(A) A word cloud of the comments



(B) Frequencies of comments by theme

FIGURE VIII

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

FIGURE VIII

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

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

V.E. Iteration

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE IV

CORRELATION BETWEEN WELL-GROOMED AND THE ALGORITHM'S PREDICTION

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE IV
CONTINUED

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

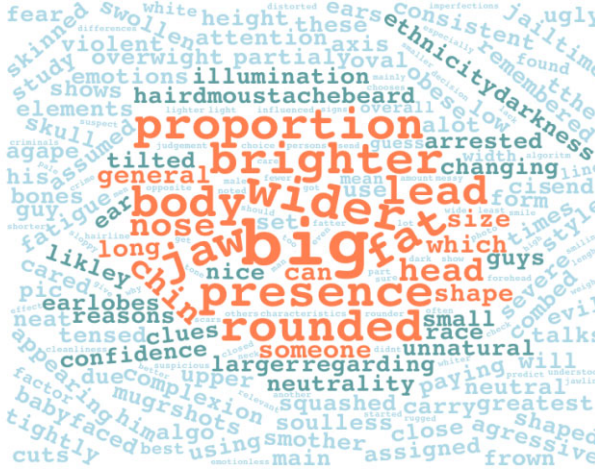
66. There are a few additional technical steps required, discussed in Online Appendix B. For details on the HIT we use to get subjects to name the new hypothesis from looking at orthogonalized morphs, and the follow-up HIT to generate independent labels for that new hypothesis or facial feature, see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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

FIGURE IX

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024



(A) A word cloud of the comments

(B) Frequencies of comments by theme

FIGURE X

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

FIGURE X

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

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

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

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

VI. EVALUATING THESE NEW HYPOTHESES

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

before turning to whether these correlations might reflect some underlying causal relationship.

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

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

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

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

VI.B. Do Practitioners Already Know This?

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

67. See Online Appendix Figure A.XIX.

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE V

CORRELATION BETWEEN HEAVY-FACED AND THE ALGORITHM'S PREDICTION

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE V
CONTINUED

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE VI

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE VI
CONTINUED

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE VII

RESULTS FROM THE CRIMINAL JUSTICE PRACTITIONER SAMPLE

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

TABLE VII
CONTINUED

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

VI.C. Exploring Causality

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

71. For additional details, see Online Appendix C.

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

VII. Conclusion

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

Supplementary Material

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

Data Availability

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

References

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz, "What We Teach about Race and Gender: Representation in Images and Text of Children's Books," *Quarterly Journal of Economics*, 138 (2023), 2225–2285. https://doi.org/10.1093/qje/qjad028

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

Berry, Diane S., and Leslie Zebrowitz-McArthur, "What's in a Face? Facial Maturity and the Attribution of Legal Responsibility," *Personality and Social Psychology Bulletin*, 14 (1988), 23–33. https://doi.org/10.1177/0146167288141003

Bertrand, Marianne, and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94 (2004), 991–1013. https://doi.org/10.1257/0002828042002561

Bjornstrom, Eileen E. S., Robert L. Kaufman, Ruth D. Peterson, and Michael D. Slater, "Race and Ethnic Representations of Lawbreakers and Victims in Crime News: A National Study of Television Coverage," *Social Problems*, 57 (2010), 269–293. https://doi.org/10.1525/sp.2010.57.2.269

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

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

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

Carleo, Giuseppe, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová, "Machine Learning and the Physical Sciences," *Reviews of Modern Physics*, 91 (2019), 045002. https://doi.org/10.1103/RevModPhys.91.045002

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

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

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

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

Dobbie, Will, Jacob Goldin, and Crystal S. Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108 (2018), 201–240. https://doi.org/10.1257/aer.20161503

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

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

Eberhardt, Jennifer L., Paul G. Davies, Valerie J. Purdie-Vaughns, and Sheri Lynn Johnson, "Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes," *Psychological Science*, 17 (2006), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

Eren, Ozkan, and Naci Mocan, "Emotional Judges and Unlucky Juveniles," *American Economic Journal: Applied Economics*, 10 (2018), 171–205. https://doi.org/10.1257/app.20160390
Frieze, Irene Hanson, Josephine E. Olson, and June Russell, "Attractiveness and Income for Men and Women in Management," *Journal of Applied Social Psychology*, 21 (1991), 1039–1057. https://doi.org/10.1111/j.1559-1816.1991.tb00458.x
Fryer, Roland G., Jr, "An Empirical Analysis of Racial Differences in Police Use of Force: A Response," *Journal of Political Economy*, 128 (2020), 4003–4008. https://doi.org/10.1086/710977
Fudenberg, Drew, and Annie Liang, "Predicting and Understanding Initial Play," *American Economic Review*, 109 (2019), 4112–4141. https://doi.org/10.1257/aer.20180654
Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 57 (2019), 535–574. https://doi.org/10.1257/jel.20181020
Ghandeharioun, Asma, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W. Picard, "DISSECT: Disentangled Simultaneous Explanations via Concept Traversals," arXiv preprint arXiv:2105.15164 2022. https://doi.org/10.48550/arXiv.2105.15164
Goldin, Claudia, and Cecilia Rouse, "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *American Economic Review*, 90 (2000), 715–741. https://doi.org/10.1257/aer.90.4.715
Goncalves, Felipe, and Steven Mello, "A Few Bad Apples? Racial Bias in Policing," *American Economic Review*, 111 (2021), 1406–1441. https://doi.org/10.1257/aer.20181607
Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, 27 (2014), 2672–2680.
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572, 2014. https://doi.org/10.48550/arXiv.1412.6572
Grogger, Jeffrey, and Greg Ridgeway, "Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness," *Journal of the American Statistical Association*, 101 (2006), 878–887. https://doi.org/10.1198/016214506000000168
Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2 (Berlin: Springer, 2009).
He, Siyu, Yin Li, Yu Feng, Shirley Ho, Siamak Ravanbakhsh, Wei Chen, and Barnabás Póczos, "Learning to Predict the Cosmological Structure Formation," *Proceedings of the National Academy of Sciences*, 116 (2019), 13825–13832. https://doi.org/10.1073/pnas.1821458116
Heckman, James J., and Burton Singer, "Abducting Economics," *American Economic Review*, 107 (2017), 298–302. https://doi.org/10.1257/aer.p20171118
Heyes, Anthony, and Soodeh Saberian, "Temperature and Decisions: Evidence from 207,000 Court Cases," *American Economic Journal: Applied Economics*, 11 (2019), 238–265. https://doi.org/10.1257/app.20170223
Hoekstra, Mark, and CarlyWill Sloan, "Does Race Matter for Police Use of Force? Evidence from 911 Calls," *American Economic Review*, 112 (2022), 827–860. https://doi.org/10.1257/aer.20201292
Hunter, Margaret, "The Persistent Problem of Colorism: Skin Tone, Status, and Inequality," *Sociology Compass*, 1 (2007), 237–254. https://doi.org/10.1111/j.1751-9020.2007.00006.x
Jordan, Michael I., and Tom M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science*, 349 (2015), 255–260. https://doi.org/10.1126/science.aaa8415
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and Anna Potapenko et al., "Highly Accurate Protein Structure

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

Korot, Edward, Nikolas Pontikos, Xiaoxuan Liu, Siegfried K. Wagner, Livia Faes, Josef Huemer, Konstantinos Balaskas, Alastair K. Denniston, Anthony Khawaja, and Pearse A. Keane, "Predicting Sex from Retinal Fundus Photographs Using Automated Deep Learning," *Scientific Reports*, 11 (2021), 10286. https://doi.org/10.1038/s41598-021-89743-x

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

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

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

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

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

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

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

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

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

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

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

Mobley, R. Keith, *An Introduction to Predictive Maintenance* (Amsterdam: Elsevier, 2002).
Mullainathan, Sendhil, and Ziad Obermeyer, "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care," *Quarterly Journal of Economics*, 137 (2022), 679–727. https://doi.org/10.1093/qje/qjab046
Mullainathan, Sendhil, and Jann Spiess, "Machine Learning: an Applied Econometric Approach," *Journal of Economic Perspectives*, 31 (2017), 87–106. https://doi.org/10.1257/jep.31.2.87
Murphy, Allan H., "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology and Climatology*, 12 (1973), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2
Nalisnick, Eric, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan, "Do Deep Generative Models Know What They Don't Know?," arXiv preprint arXiv:1810.09136, 2018. https://doi.org/10.48550/arXiv.1810.09136
Narayanaswamy, Arunachalam, Subhashini Venugopalan, Dale R. Webster, Lily Peng, Greg S. Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C. Nelson, and Avinash V. Varadarajan, "Scientific Discovery by Generating Counterfactuals Using Image Translation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention,* (Berlin: Springer, 2020), 273–283. https://doi.org/10.1007/978-3-030-59710-8_27
Neumark, David, Ian Burn, and Patrick Button, "Experimental Age Discrimination Evidence and the Heckman Critique," *American Economic Review*, 106 (2016), 303–308. https://doi.org/10.1257/aer.p20161008
Norouzzadeh, Mohammad Sadegh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune, "Automatically Identifying, Counting, and Describing Wild Animals in Camera-Trap Images with Deep Learning," *Proceedings of the National Academy of Sciences*, 115 (2018), E5716–E5725. https://doi.org/10.1073/pnas.1719367115
Oosterhof, Nikolaas N., and Alexander Todorov, "The Functional Basis of Face Evaluation," *Proceedings of the National Academy of Sciences*, 105 (2008), 11087–11092. https://doi.org/10.1073/pnas.0805664105
Peterson, Joshua C., David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths, "Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-Making," *Science*, 372 (2021), 1209–1214. https://doi.org/10.1126/science.abe2629
Pierson, Emma, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer, "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations," *Nature Medicine*, 27 (2021), 136–140. https://doi.org/10.1038/s41591-020-01192-7
Pion-Tonachini, Luca, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W. Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, and Benjamin Nachman et al."Learning from Learning Machines: A New Generation of AI Technology to Meet the Needs of Science," arXiv preprint arXiv:2111.13786, 2021. https://doi.org/10.48550/arXiv.2111.13786
Popper, Karl, *The Logic of Scientific Discovery* (London: Routledge, 2nd ed. 2002). https://doi.org/10.4324/9780203994627
Pronin, Emily, "The Introspection Illusion," *Advances in Experimental Social Psychology*, 41 (2009), 1–67. https://doi.org/10.1016/S0065-2601(08)00401-2
Ramachandram, Dhanesh, and Graham W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," *IEEE Signal Processing Magazine*, 34 (2017), 96–108. https://doi.org/10.1109/MSP.2017.2738401
Rambachan, Ashesh, "Identifying Prediction Mistakes in Observational Data," Harvard University Working Paper, 2021. www.nber.org/system/files/chapters/c14777/c14777.pdf

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024

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

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

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

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

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

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

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

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

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

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing Properties of Neural Networks," arXiv preprint arXiv:1312.6199, 2013. https://doi.org/10.48550/arXiv.1312.6199

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

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

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

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

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

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

© The Author(s) 2024. Published by Oxford University Press on behalf of the President and Fellows of Harvard College. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/qje/advance-article/doi/10.1093/qje/qjad055/7515309 by Imperial College London Library user on 18 March 2024