Decision Making with Quantized Priors Leads to Discrimination

Lav R. Varshney, *Senior Member, IEEE* and Kush R. Varshney, *Senior Member, IEEE*

***Abstract*—Racial discrimination in decision-making scenarios such as police arrests appears to be a violation of expected utility theory. Drawing on results from the science of information, we discuss an information-based model of signal detection over a population that generates such behavior as an alternative explanation to taste-based discrimination by the decision maker or differences among the racial populations. This model uses the decision rule that maximizes expected utility—the likelihood ratio test—but constrains the precision of the threshold to a small discrete set. The precision constraint follows from both bounded rationality in human recollection and finite training data for estimating priors. When combined with social aspects of human decision making and precautionary cost settings, the model predicts the own-race bias that has been observed in several econometric studies.**

***Index Terms*—beliefs, decision making, quantization, racial discrimination**

I. Introduction

Quantization is prevalent in economic phenomena, whether declaring recessions in macroeconomics, grouping citizens into districts in social choice theory, grading agricultural commodities in trade [1, pp. 132–135], penny shaving and similar arbitrage in finance (cf. [2]), advertisement targeting in marketing, and even resource allocation in basic microeconomics [3, p. 35]. Yet, studies of these phenomena do not draw on insights from the science of information, where compression and quantization as responses to information constraints are understood deeply [4]–[6]. The emergence of rational inattention theory [7] brings Shannon-theoretic principles into economics, but constraining mutual information implicitly assumes the asymptotic latency or system size needed for coding theorems to endow this constraint with operational significance [8].

Quantization theory [4], [5], however, is eminently applicable when latency and size are not large. It is useful for understanding the economic settings mentioned above and beyond. Here, we use a quantization-theoretic approach to develop a bounded rationality model of human decision making that may provide insight into a most troubling aspect of civic life: *racial discrimination*. In particular, our informational theory studies ensembles of decision problems people face in their daily lives when they have limited memory to characterize the specific decision problems drawn from the ensembles.

L. R. Varshney is with the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA (e-mail: varshney@illinois.edu).

K. R. Varshney is with the Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 10598 USA (e-mail: krvarshn@us.ibm.com).

The theory builds on our past work on decision-making under memory constraints [9]–[12], but further using facts from social cognition and social segregation. We note that although many of the decision-theoretic and quantization-theoretic aspects of the model were developed in [9] (together with some basic social implications), the central contribution of the present paper is in placing model implications within the social science literature by making strong connections to extant work in several fields including economics, sociology, psychology, and law. Doing so requires new mathematical results, e.g. in learning theory, which are also given.

Racial discrimination often manifests in decision-making settings where the subject and object of decision making are members of different races. This phenomenon is often attributed to a taste for discrimination where the decision maker derives utility from differential treatment of races or to statistical differences between races leading to improved decision-making performance via differential treatment. Often there is reason to suspect these effects are not a complete explanation. We ask whether there are alternative mathematical models of behavior that may also provide explanation; we give a model of limited human decision making that generates racial discrimination independent of the other two effects.

Decision making under risk and uncertainty is one of the most basic economic actions. Typical decisions include the binary choice of whether or not to hire a worker, and whether or not a rule is being violated. Such binary choice is mathematically equivalent to detecting a signal in noise (observations are never perfect). Indeed, signal detection theory forms the core of psychological models of cognition, including choice and classification [13]. Many economic analyses of choice use the decision rule that maximizes expected utility [14]–[17], the likelihood ratio test (LRT) [18]. The key parameter of the LRT is the threshold. When an optimal strategy is employed, it is a function of the prior probabilities of the two choices as well as the costs assigned to the two types of errors.

If several decision problems with differing prior probabilities are to be solved, the optimal strategy is to also set thresholds differently each time. Examples of such scenarios include calling of fouls by sports referees and making of arrests by police officers, as each player/citizen has a different prior probability of committing a foul or crime. Now consider a decision maker who categorizes or groups together population members by prior probability, e.g. a police officer using categories *law-abiding*, *delinquent*, *criminal*, and *nefarious* in the decision to arrest a citizen instead of using the citizen's precise individual prior probability of committing a crime.

We describe a model of decision making under uncertainty that incorporates population categorization (and consequently limited threshold precision), mathematized using quantization theory [4], [9]. The proposed model aims to explain the bias based on race, the so-called racial profiling observed in several decision-making scenarios of the type described, including foul calls by National Basketball Association (NBA) referees, arrests for minor offences by police, and searches of stopped vehicles by police [19]–[21].

LRTs are not only optimal detection rules, but psychology experiments suggest human decision makers also employ them [22], [23]. Psychology experiments also suggest humans are able to use prior probabilities in decision rules when available in natural formats [24]. The use of prior probabilities to set the threshold makes the decision rule Bayesian; Bayesian models are common in both psychology and economics [25], [26].

There are two motivations for considering decision makers who categorize according to prior probabilities. First, using a different threshold for different decisions puts much strain on a human decision maker (the referee or police officer). When decision makers make decisions on members of a population, information processing constraints lead to categorical and coarse thinking [27]–[29]. Second, if the prior probabilities are learned by the decision maker from noisy samples, then categorization helps prevent the statistical phenomenon of overfitting [30, Ch. 5.3].[1]

To reach an information-based model that generates discrimination, rather than a taste-based, statistical, or implicit one [32]–[34][2] we require a few ingredients in addition to limited precision in the LRT threshold. These additional ingredients arise from social structure [35] and social cognition [36].

Distinct from the categorization due to information processing limits or generalizable learning, social cognition theory predicts another form of categorization. There is evidence that people tend to automatically categorize others according to race in multiracial societies [37]–[39]. As a consequence, when decision makers make decisions on members of a racially mixed population, they may use different decision rules for different racial groups.

The social interaction patterns of decision makers may also influence their cognitive processes. One might think social interaction history plays no role in economic decisions, but social life and economic life are inextricably intertwined. We assume human decision makers are unable to completely discount social experience in determining how to deploy their limited decision-making resources (spillover may be due to the use of common brain regions [40]). In studying racial bias, the relevant parameter of social interaction is race. Due to segregation in social life, there is greater intra-race interaction than inter-race interaction, see e.g. [41], [42].

In our model, the mismatch between exposure to different racial populations in social life and in economic life is a central cause for the social welfare loss that may accompany racial bias in decision making. We call this social welfare loss the *price of segregation*. The price of segregation is a purely informational phenomenon rather than one involving social capital or in-group/out-group trust issues [43]–[45].

In most decision-making scenarios, one of the hypotheses leads to no action and is the default (null hypothesis), whereas the other requires action (alternative hypothesis). For example, no foul call, no arrest, and no search are default actions. We call the utility function of the decision maker *precautionary* if the cost of deciding for the active hypothesis when the default hypothesis is true is much greater than the cost of deciding for the default hypothesis when the active hypothesis is true. We call the utility function *dauntless* if the cost ordering is reversed.

In the overall model we develop, the utility function of the decision maker must be precautionary to explain greater active outcomes for the population that is of different race than the decision maker, such as those found in many empirical observations. If the utility function is dauntless, then the model predicts greater active outcomes for the population that is of the same race, which has also been observed occasionally, e.g. [46]. The proposed model generates the interesting phenomenon that the decision-making utility function of the decision maker has a fundamental effect on the nature of racial discrimination, a phenomenon that has not been described before by a single model.

II. Related Prior Work

The model we propose is based on categorization, specifically the quantization of prior probabilities, due to memory constraints and social cognition factors, and generates discrimination in decision making. Here, we review prior work on both quantization and on discrimination.

A. Bounded Rationality and Quantization

Of late, economic theories and models have started taking bounded rationality into account [47]–[50]. Two basic kinds of models of imperfect memory may be considered. In the first kind, *truly bounded rationality* [49], there is an explicit model of memory in which the decision maker is not aware of the limitation: the goal then is to look for the implications of the constraints [51], [52]. In the second kind of model, *costly rationality* [49], the decision maker has a memory constraint but is fully aware of this limitation. The decision maker can then use an optimal strategy under the constraint [53]–[56]. Our model follows the costly rationality paradigm.

Although optimal quantization theory [4], [5] has not appeared in previous economic models, forms of categorization have. Dow looks at a sequential decision making scenario, where the decision maker is deciding whether to purchase from one vendor or another [53]. The decision maker first observes the price of the first vendor, but does not purchase. Then, the decision maker observes the price of the second vendor,

[1] A passage in [31] summarizes this perspective: "One obvious explanation would be along the lines of bounded rationality — due to limitations of the human mind, simpler theories are easier to conceive of, recall, communicate, and test. While this explanation is certainly valid, it is worth our while to ask whether there might be an innate advantage in preferring simpler theories. That is, assume that you are programming an organism endowed with unlimited computational power. When the organism compares different theories, would you like to instruct the organism to prefer simpler ones (keeping accuracy and generality constant)? One may attempt an affirmative answer along the following lines. . . . [The] argument is that the preference for simplicity is a guarantee against overfitting."

[2] See [32] for discussion on various forms of discrimination.

compares the prices, and purchases from the lower-priced vendor. However, due to bounded rationality, the decision maker only remembers a quantized version of the first price when comparing. The problem setup in that work is different than the scenario we discuss, but the analysis has similar flavor.

Mullainathan also considers a sequential scenario and is concerned with learning beliefs from data, but decision making is not a part of the framework [57]. In a sequence of observations, the decision maker perfectly knows the state of an object. If a rational decision maker keeps making observations for a long time, the empirical frequencies of the observations converge to the true probabilities. The model, however, inserts bounded rationality into the learning process. Quantization regions partition the probability simplex; sequential updates to the probabilities are based not on likelihoods from data, but quantized versions of likelihoods from data. Because decision making is not a part of the picture in the work, the optimization criterion is not correctly matched; general learning and learning for a particular purpose such as decision making are not always equivalent.

B. Discrimination

Our model falls under the economics of information [58] and is a theory of information-based discrimination, but is distinct from statistical discrimination [33] and implicit discrimination [34]. Quantization serves to reduce the available amount of information. Unlike other studies of discrimination, we assume that the racial populations are identical in the relevant quality, that the measurement process for judging quality is also identical, and that there are no dynamic effects.

A possible way to model discrimination is that it happens because blacks commit more fouls or whites commit more crimes. This assumption that the propensity to commit a foul or crime is different in different racial populations is called statistical discrimination [33]. For example, [59] assumes the two populations are not identical in propensity and [60] assumes population membership and propensity are correlated. These heterogeneity assumptions are not required in the model proposed here; we specifically focus on the situation when the populations have identical distributions of prior probabilities.

Another proposed explanation for discrimination is via a dynamic process in which a group fails to invest in human capital because it is not valued by the decision maker, a self-fulfilling prophecy [33], [61], [62]. Knowles, et al. similarly argue that non-identical propensities arise as an equilibrium in the interaction of police officers and citizens, also proposing an empirical test between statistical discrimination and taste-based discrimination [63]; Anwar et al. give an alternate model and test that makes use of the police officer's race [64]. Our model does not require a dynamic equilibrium but generates racial discrimination in a static setting of identical prior probability distributions.

Another potential modeling component is that population members of a different race as the decision maker are observed with greater uncertainty than population members of the same race. In addition to non-identical populations, [59] assumes there are different levels of noise in measuring different populations. Other work enforce identicalities over the populations, but requires some difference in estimator performance to explain discrimination [65], [66]. Similarly, [67] works with identical quality distributions in the two populations, but assumes a different quality of measurements depending on whether the decision maker and applicant for employment are of the same or different race, specifically in tournament situations of screening; implications of their model are fairly similar to what we find. Unlike these works, our model uses identical measurement quality for different populations.

Like the model developed herein, Fryer and Jackson [28] do not require differences among populations or measurements, or feedback to explain bias. Their work discusses how decision makers use categorization for information processing, how decisions for minority populations may be less accurate, and how this may lead to discrimination against minority groups even without malevolent intent. However, the model faces shortcomings introduced by not considering the race of the decision maker. It also implicitly requires the objects of decision making to be members of minority groups, rather than being a general model. Further, [28] does not consider categorization specifically for decision making and yet despite this simplification says "we are unable to prove general results about optimal categorizations, and this failure is itself informative" going on to say that "providing a full description of what an optimal categorization looks like . . . is a hard problem." The science of information overcomes both of these shortcomings [4]. Unlike the Fryer–Jackson model, our explicit decision-theoretic model implies that decision-making attitude fundamentally impacts the nature of racial discrimination. Additionally, our model yields the price of segregation as a measure of social welfare loss, something wholly absent in [28]. Finally, our model has stronger empirical content, as it yields a quantitation of difference in differences in decision-making performance.

III. Decision-Making Model

In this section, we propose a model of decision making where the prior probabilities that go into the LRT threshold are optimally quantized, motivated by recall limitations of people or generalizable learning. (The learning motivation is explicated in Section IV.) The model is further extended to have separate optimal quantizers for different racial populations, as motivated by social cognition factors.

A. Bayes Risk and Likelihood Ratio Test Detection Rule

Consider signal detection in which a decision maker uses a noisy observation Y to determine whether an object (citizen or player) is in state h_0 or h_1. State h_0 corresponds to a null hypothesis such as no foul committed, whereas h_1 corresponds to an alternative hypothesis such as foul committed. Noisy observations on whether or not a foul was committed are modeled by likelihood functions $f_{Y|H}(y|h_0)$ and $f_{Y|H}(y|h_1)$. The object has prior probability p_0 of being in state h_0 and $p_1 = 1 - p_0$ of being in state h_1, i.e. $p_0 = \Pr[H = h_0]$ and $p_1 = \Pr[H = h_1]$. In a population of objects, each object may have a different prior probability. That is, different citizens may have different prior propensities for crime. The population is modeled by a probability density function $f_{P_0}(p_0)$ supported on the unit interval (two-dimensional probability simplex); this is a probability distribution on probabilities.

The detection rule $\hat{h}(y)$ of the decision maker is the LRT:

$$\frac{f_{Y|H}(y|h_1)}{f_{Y|H}(y|h_0)} \underset{\hat{h}(y)=h_1}{\overset{\hat{h}(y)=h_0}{\lessgtr}} \frac{c_{10}a}{c_{01}(1-a)}, \tag{1}$$

where c_{ij} is the non-negative cost of deciding h_j when the true state is h_i (we assume the decision maker assigns no cost to correct decisions). The parameter a weights the decision rule so as to allow the incorporation of prior beliefs. There are two types of errors, with the following probabilities:

$$p_E^{\mathrm{I}} = \Pr[\hat{h}(Y) = h_1 | H = h_0],$$
$$p_E^{\mathrm{II}} = \Pr[\hat{h}(Y) = h_0 | H = h_1].$$

The Bayes risk, the performance of the decision rule, may be expressed in terms of those error probabilities as:

$$J(p_0, a) = c_{10} p_0 p_E^{\mathrm{I}}(a) + c_{01}(1-p_0) p_E^{\mathrm{II}}(a). \tag{2}$$

Error probabilities depend on a through $\hat{h}(\cdot)$, given in (1). If the parameter a is set so $a = p_0$, the decision rule (1) is the Bayes optimal decision rule, minimizing Bayes risk (2). The function of one variable $J(p_0, p_0)$ is zero at the points $p_0 = 0$ and $p_0 = 1$ and is positive-valued, strictly concave, and continuous in the interval $(0, 1)$. Notice that $J(p_0, a)$ is a linear function of p_0 with slope $(c_{10}p_E^{\mathrm{I}}(a) - c_{01}p_E^{\mathrm{II}}(a))$ and intercept $c_{01}p_E^{\mathrm{II}}(a)$; furthermore $J(p_0, a)$ is tangent to $J(p_0, p_0)$ at a.

B. Quantized Prior

The choice $a = p_0$ in the LRT (1) is the fully rational one. An essential piece of our model, however, is that the decision maker is bounded to be a coarse thinker and must use the same prior belief parameter a for different objects. Thus the decision maker only has access to which category an object belongs to, when making decisions about that object. Categorization of objects (in the prior probability space) is modeled as a quantizer for the population distribution $f_{P_0}(p_0)$. A K-point quantizer of $f_{P_0}(p_0)$ partitions the interval $[0, 1]$ into K regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$. For each quantization region $\mathcal{R}_k$, there is a representation point a_k to which elements are mapped. This value a_k may be thought of as the prior probability for a prototype member of the kth category. For deterministic regular quantizers,[3] regions are subintervals $\mathcal{R}_1 = [0, b_1]$, $\mathcal{R}_2 = (b_1, b_2]$, ..., $\mathcal{R}_K = (b_{K-1}, 1]$ and representation points a_k are in $\mathcal{R}_k$. A quantizer can be viewed as a nonlinear function $q_K(\cdot)$ such that $q_K(p_0) = a_k$ for $p_0 \in \mathcal{R}_k$.

With constraint, the decision maker uses the prior belief parameter $a = q_K(p_0)$ in the LRT (1). There are many possible quantization functions or categorizations of objects; following costly rationality, $q_K(\cdot)$ should be optimal in terms of decision-making performance. We propose that $q_K(\cdot)$ (for fixed K) minimize $D = E[J(P_0, q_K(P_0)) - J(P_0, P_0)]$, where the expectation is with respect to $f_{P_0}(p_0)$. As such, the quantization fidelity criterion is the difference between the Bayes risk with quantized priors and the optimal Bayes risk with unquantized priors.

We term the quantization distortion function $d(p_0, a) = J(p_0, a) - J(p_0, p_0)$ the *Bayes risk error* (BRE). As the difference of the tangent line to a strictly convex function and that convex function, the BRE is a Bregman divergence [70]. It is non-negative and only equal to zero when $p_0 = a$, is continuous and strictly convex as a function of $p_0 \in (0, 1)$ for all a, and for any deterministic LRT has exactly one stationary point as a function of $a \in (0, 1)$ for all p_0, which is a minimum. The BRE is quasiconvex as a function of $a \in (0, 1)$ for all p_0: a slightly weaker condition than having exactly one stationary point that is a minimum [71, Section 3.4.2].

1) Quantized Prior Mathematical Details: In general the design of an optimal quantizer does not have a closed-form solution. Nevertheless, there are three conditions that an optimal quantizer must satisfy. These three necessary conditions are known as the centroid condition, nearest neighbor condition, and the zero probability of boundary condition [4], [72]–[74].

Let us define the centroid $\mathrm{cent}(\mathcal{R})$ of a random variable P_0 in a region $\mathcal{R}$ with respect to a distortion function $d(\cdot, \cdot)$ as

$$\mathrm{cent}(\mathcal{R}) = \arg\min_a E[d(P_0, a) \mid P_0 \in \mathcal{R}]. \tag{3}$$

Condition 1 (Centroid): For a given set of quantization regions $\{\mathcal{R}_k\}$, the optimal representation points satisfy $a_k = \mathrm{cent}(\mathcal{R}_k)$ and for BRE (and any other Bregman divergence [75]), $a_k = E[P_0 \mid P_0 \in \mathcal{R}]$.

Condition 2 (Nearest Neighbor): For a given set of representation points $\{a_k\}$, the quantization regions satisfy

$$\mathcal{R}_k \subset \{p_0 \mid d(p_0, a_k) \le d(p_0, a_j) \text{ for all } j \ne k\}. \tag{4}$$

For BRE, the boundary point $b_k \in [a_k, a_{k+1}]$ is the abscissa of the point at which the lines $J(p_0, a_k)$ and $J(p_0, a_{k+1})$ intersect. This point is

$$b_k = \frac{c_{01}\left(p_E^{\mathrm{II}}(a_{k+1}) - p_E^{\mathrm{II}}(a_k)\right)}{c_{01}\left(p_E^{\mathrm{II}}(a_{k+1}) - p_E^{\mathrm{II}}(a_k)\right) - c_{10}\left(p_E^{\mathrm{I}}(a_{k+1}) - p_E^{\mathrm{I}}(a_k)\right)}.$$

The intersection point is obtained by manipulating the slopes and intercepts of $J(p_0, a_k)$ and $J(p_0, a_{k+1})$.

Condition 3 (Zero Probability Boundary): The third necessary condition for quantizer optimality only arises when dealing with probability distributions that contain a discrete component and are thus not absolutely continuous. The random variable P_0 must have zero probability of occurring at a boundary between quantization regions. This condition from optimal quantization theory [4] was rediscovered in economics by Fryer and Jackson [28, Lem. 1].

Theorem 1 ([4], [72]–[74]): Conds. 1, 2, and 3 are necessary for a quantizer to be optimal.

If additional conditions are met, then the necessary conditions for optimality are also sufficient for local optimality.

Theorem 2: If the following conditions hold:

1) $f_{P_0}(p_0)$ is positive and continuous in $(0, 1)$;

[3]There is no loss of optimality in restricting attention to deterministic regular quantizers and the Voronoi partitions they imply since the Bayes risk error $d(p_0, a)$ is strictly convex in p_0 for all a, see [4, Lem. 6.2.1]. Moreover according to the psychological principles of grouping according to proximity, humans are thought to categorize according to considerations of perceived similarity between objects [68], [69], which is regular quantization.

2) $\int_0^1 d(p_0, a) f_{P_0}(p_0) dp_0$ is finite for all a; and
3) $d(p_0, a)$ is zero only for $p_0 = a$, is continuous in p_0 for all a, and is continuous and quasiconvex in a,

then the nearest neighbor condition, centroid condition, and zero probability of boundary conditions are sufficient to guarantee local optimality of a quantizer.

Proof: Minor modification of results in [76]. ■

One can note that the first and second conditions of Thm. 2 are met by common distributions such as the family of beta distributions. The third condition is satisfied by BRE, as described above. The conditions for optimality suggest the iterative Lloyd–Max algorithm [4], [73], [74].

C. Separate Quantizers for Different Racial Populations

We now turn to the situation in which the decision maker must deal with subpopulations distinguished according to a socially observable part of identity like race [77]. For clarity and ease of connection to empirical studies, we restrict to two groups and use 'black' and 'white' to denote them. The rational coarse-thinking decision maker should ignore the irrelevant dimension of race altogether and simply partition along the p_0 dimension, but social cognition constraints prevent the decision maker from doing so.

Automaticity of racial categorization results in two quantizers designed separately for the two populations. The total quota on representation points, K_t, is split into some number of points for whites and some number for blacks, denoted $K_t = K_w + K_b$. The separate quantizers may then be denoted $q_{K_w}(\cdot)$ and $q_{K_b}(\cdot)$.

We can extend the definition of mean BRE to two populations as:

$$D^{(2)} = \tfrac{m_w}{m_w+m_b} E[J(P_0, q_{K_w}(P_0))] + \tfrac{m_b}{m_w+m_b} E[J(P_0, q_{K_b}(P_0))] - E[J(P_0)], \quad (5)$$

where m_w and m_b are the number of whites and blacks within the social and economic ambit of the decision maker. Under costly rationality, the goal is to minimize this extended BRE by finding optimal quantizers $q_{K_w}(\cdot)$ and $q_{K_b}(\cdot)$ and optimal allocation of representation points K_w and K_b.

The model we propose assumes that the two populations are identical. Thus $q_{K_w}(\cdot)$ and $q_{K_b}(\cdot)$ should be designed as discussed earlier. The problem reduces to minimizing expected BRE over all $K_t - 1$ possible allocations of K_w and K_b. Although there are sophisticated algorithms for optimal allocation of levels [78], just measuring the performance of all allocations and choosing the best one suffices.

Fryer and Jackson [28] previously suggested it is better to allocate more representation points to a majority population than to a minority population. Optimal allocation in this model yields the same result when the notion of majority and minority are with respect to the decision maker's interaction pattern. If m_w is larger than m_b, it is better to allocate more representation points to whites whereas if m_b is larger than m_w, it is better to allocate more representation points to blacks. An example of optimal allocation is shown in Fig. 1.

In this section, we have proposed a Bayesian LRT that incorporates prior probability quantization to model human decision making. The model has been further extended for decision making on distinct racial populations through separate quantization functions to incorporate social cognition factors. Section V quantitatively and qualitatively shows the implications of this model.

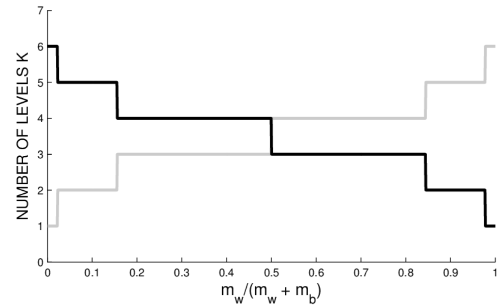


Fig. 1. Optimal allocation of quantizer sizes to the white population and black population for $K_t = 7$ as a function of the proportion of whites. The gray line is K_w and the black line is K_b. The distribution of the prior probability is beta(5,2), the measurements are observed through additive white Gaussian noise with unit signal to noise ratio, and the Bayes costs are $c_{10} = c_{01} = 1$.

IV. Learning

To this point in the paper, the motivation for the LRT model with quantized priors has been bounded rationality on the part of the human decision maker. In this section, we present an alternate motivation that is statistical in nature—the prevention of overfitting—and arises endogenously without appealing to human memory or information processing constraints. Overfitting is the statistical phenomenon of an overly complex model learned from training data having poor predictive performance or generalization on new unseen data drawn from the same distribution as the training data [79].

A. Decision Maker Training Model

The prior probability quantization described thus far in the paper is based on a known population distribution $f_{P_0}(p_0)$. Here we consider the situation in which this distribution is assumed to be unknown directly. Available to the decision maker is a set of noisy estimates of prior probabilities $\hat{p}_{0,j}$, $j = 1, \ldots, n$, i.e. noisy prior probably estimates for n players or citizens in the population. These n players are drawn independently from $f_{P_0}(p_0)$ and have true prior probabilities $p_{0,j}$, $j = 1, \ldots, n$.

The noisy estimates are specifically as follows. For a particular player j, the referee observes m_j plays of that player. A play is of hypothesis h_0 with probability $p_{0,j}$ and hypothesis h_1 with probability $1 - p_{0,j}$. The referee observation is a random variable Y_j which has mean zero when the hypothesis is h_0 and has mean one when the hypothesis is h_1; the distribution around the specified mean is identical under each hypothesis. A simple unbiased, consistent, universal estimate for $p_{0,j}$ from these m_j samples is

$$\hat{p}_{0,j} = 1 - \frac{1}{m_j} \sum_{i=1}^{m_j} y_{j,i}, \quad (6)$$

which we assume the decision maker takes as the prior probability for player or citizen j.

The proposed model of samples may be understood by considering what data the decision maker might observe when learning about the population. Populations contain a finite number of objects, which is why the finite number n is included in the model. In training, the decision maker can only observe each object a finite number of times m_j. When the decision maker is learning about the population, perfect measurements of the object state or hypothesis may not be available. These measurements will generally be noisy; the variables $y_{j,i}$ capture any measurement noise. Overall, this formulation models a referee who learns prior probabilities of players from repeated noisy observations of their plays.

The estimate $\hat{p}_{0,j}$ may be used by the decision maker directly in the LRT threshold for citizen j. Alternatively, the collection of $\hat{p}_{0,j}$ may first be grouped into $K < n$ clusters with the threshold for citizen j a function of his or her cluster center. The appropriate clustering here is K-means clustering to minimize BRE [80]. Such clustering is a way to reduce complexity and prevent overfitting. Note that clustering with $K = n$ is equivalent to not clustering.

As the number of samples n increases, the sequence of clusterings learned from $p_{0,1}, \ldots, p_{0,n}$ converges to the quantizer designed from $f_{P_0}(p_0)$ under conditions on the distortion function met by the BRE due to convexity and quasiconvexity in its two arguments [81], [82]. In our case it is not the $p_{0,1}, \ldots, p_{0,n}$, but the consistent estimates $\hat{p}_{0,1}, \ldots, \hat{p}_{0,n}$ that are available, so all of the m_j must also increase for overall convergence. However, our interest is not in studying the situation of n and m_j growing without bound, but to consider the situation with finite n and m_j.

B. Decision Maker Generalization Behavior

With prior probability estimation, minimum mean BRE K-means clustering, and LRT detection as defined above, we would like to examine the behavior of the minimum mean BRE as a function of K, which tells us the generalizability of the detection rule learned from m_j noisy samples.

For specificity, we focus on a model in which the likelihood functions $f_{Y|H}(y|h_0)$ and $f_{Y|H}(y|h_1)$ are Gaussian with means zero and one, and the same variance. We take the noisy measurements that the decision maker uses in learning about the population to be a mixture of two Gaussians with means zero and one, and the same variance. The mixture weights are $p_{0,j}$ and $1 - p_{0,j}$. Thus the noisy measurements used by the decision maker in learning about the population are of the same type as used when doing the LRT. The number of observations per player or citizen, m_j, is taken to be the same value m across all n players or citizens in the population.

The mean BRE of this formulation with noisily learned prior probabilities is plotted as a function of the number of clusters K for a particular choice of parameters. (The same behavior repeats for other choices of parameters.) A zoomed in portion appears in Fig. 3. The black lines are the mean BRE values of the formulation with learning, and the gray lines are the mean BRE values of the formulation with quantization of known

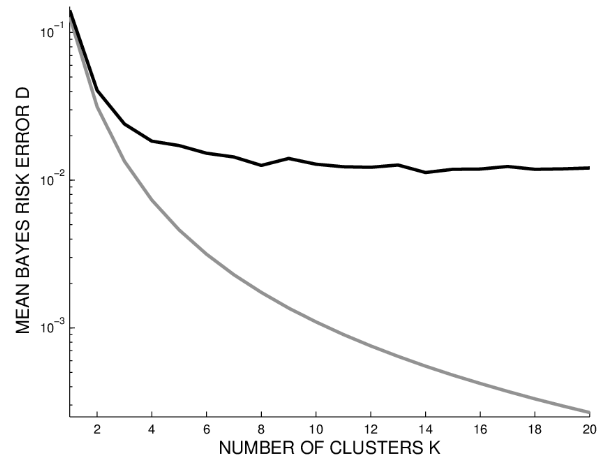


Fig. 2. Mean BRE with quantization of known $f_{P_0}(p_0)$ (gray line), and with K-means clustering of noisy prior probability estimates with $m = 100$ (black line), uniform P_0, and $c_{10} = 1$, $c_{01} = 4$.

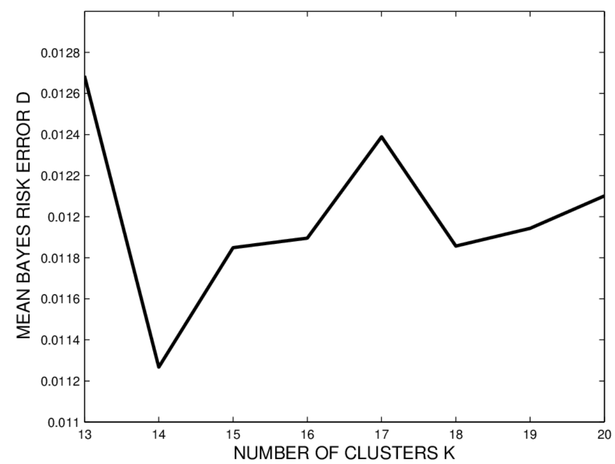


Fig. 3. Zoomed in portion of Fig. 2.

$f_{P_0}(p_0)$. The black lines are not smooth due to the stochastic nature of the random sampling and the discrete nature of K-means clustering.

The mean BRE of clustering noisy estimates is greater than the mean BRE of quantizing the population distribution. Importantly, whereas the mean BRE of quantization goes to zero as K goes to infinity, the mean BRE of clustering does not go to zero. In fact, there is an intermediate value of K at which the mean BRE is minimized. This is most obviously seen in Fig. 3. This behavior is a manifestation of the overfitting phenomenon. By limiting complexity through limiting prior probability precision, decision-making performance improves.

We further examine the effect of the training set size. The best value of K for a given m is plotted in Fig. 4. Just as importantly, we see that a larger training set implies that more clusters should be used. Since the estimate $\hat{p}_{0,j}$ gets closer to $p_{0,j}$ as m increases, less regularization in the form of clustering is needed.

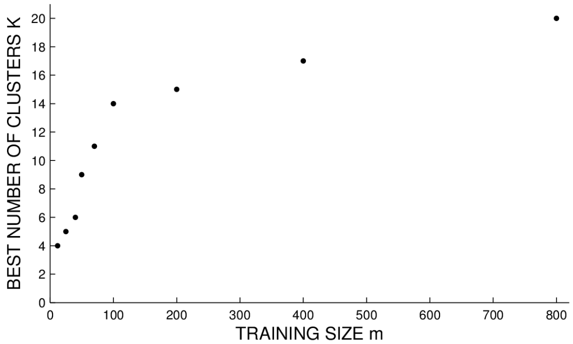


Fig. 4. Number of clusters that minimize mean BRE with K-means clustering of noisy prior probability estimates for several values of m, uniform P_0, and $c_{10} = 1$, $c_{01} = 4$.

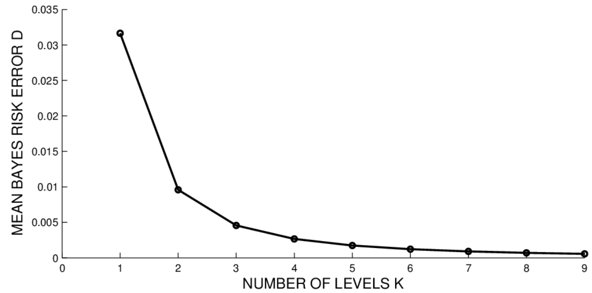


Fig. 5. Mean BRE of optimal quantizers as a function of the number of quantizer cells K. The distribution of the prior probability is beta(5,2), the measurements are observed through additive white Gaussian noise with unit signal to noise ratio, and the Bayes costs are $c_{10} = c_{01} = 1$.

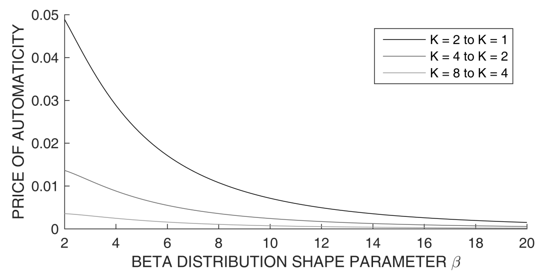


Fig. 6. Price of automaticity (mean BRE difference of optimal quantizers in halving the number of levels), as a function of the concentration parameter β for the distribution of the prior probability being beta(2,β), with the measurements observed through additive white Gaussian noise with unit signal to noise ratio, and Bayes costs $c_{10} = c_{01} = 1$.

The formulation of this section exhibits the hallmark behavior of overfitting in statistical learning. This behavior may be analyzed theoretically within the context of the class proportion estimation (CPE) problem [83]. For example, existing risk bounds for the CPE problem (in which complexity is controlled by other means) [84]–[86] may be modified by characterizing the Rademacher complexity of K-means clustering via ϵ-entropy [87].

To summarize, even when quantization/clustering is not imposed as a form of bounded rationality on the part of the decision maker, we discover that clustering is optimal from a statistical learning perspective. Furthermore, it is true that if we have two different amounts of training m_w and m_b, then $K_w^* \geq K_b^*$ for $m_w > m_b$, and $K_b^* \geq K_w^*$ for $m_b > m_w$. These relationships are the same as before, but here obtained without decision maker processing limitations.

V. Model Predictions

Having established a mathematical model of optimal decision making under memory and social cognition constraints, we explicitly put forth some model predictions.

A. *Price of Automaticity*

Bayes risk performance does not get worse when the number of quantization levels K is increased, under optimal memory-constrained decision making. Let

$$D^*(K) = \sum_{k=1}^{K} \int_{\mathcal{R}_k^*} d(p_0, a_k^*) f_{P_0}(p_0) dp_0$$

denote mean BRE for an optimal K-point quantizer. Then $D^*(K)$ is a non-increasing sequence of K [9]. In typical settings (Fig. 5), performance strictly improves with each increase in the number of quantization levels.

A referee will perform better with more categories rather than fewer. A police officer confronting an individual with whom she has prior experience will make a better decision if she has the mental categories *law-abiding*, *delinquent*, *criminal*, and *nefarious*, rather than just *good* and *bad*.

Since there is a finite budget of categories, automaticity of racial categorization leads to coarser categories and loss in BRE performance for each of the subpopulations, the *price of automaticity*. In Fig. 5, halving the number of levels (e.g. in evenly splitting across two subpopulations) from 8 to 4 has much less effect on mean BRE than going from 4 to 2 or 2 to 1. As the number of quantization levels tends to infinity, halving has negligible effect (see [9, Sec. IV] for precise high-resolution characterization). Fig. 6 demonstrates a decrease in the price of automaticity as the prior probability distribution becomes more concentrated. If the prior distribution were supported on a finite set of mass points with cardinality less than K_w and K_b, the price of automaticity would be zero.

If contrary to our assumptions that subpopulations have identical prior probability distributions, there are significant statistical differences among subpopulations (as in statistical discrimination [33]), separate quantizers with fewer levels may yet be better. This is if the performance degradation from mismatch is more than performance degradation from rate loss.

B. *Difference in Differences*

Many aspects of social segregation might impact decision making, but our interest is in exposure: the amount of contact and interaction between members of different groups [35], [41]. Due to racial isolation from social segregation, there is greater intra-population interaction than inter-population interaction. Whites interact more with whites whereas blacks interact more with blacks. This is how even if economic life is identical, small *economically-irrelevant* differences in cognitively intertwined social life can have an impact.

In the model, one would expect $m_w/(m_w + m_b)$ of a white decision maker to be greater than $m_w/(m_w + m_b)$ of a black decision maker. Putting this fact together with optimal representation point allocation and monotonicity implies a white decision maker would perform worse than a black decision maker when dealing with blacks and a black decision maker would perform worse than a white decision maker when dealing with whites, measuring by expected Bayes risk.

Econometric studies provide a source for comparison to the proposed decision making model. A major difficulty in interpreting these studies, however, is that ground truth is not known. Higher rates of arrest or foul calls may be explained by either a greater p_E^{I} or smaller p_E^{II}. It is possible that a greater probability of missed fouls would actually decrease the number of fouls called. This motivates a closer look at the Bayes risk; we tease it apart into its constituent parts and examine the Bayes costs in detail.

Using sports officiating as our running example, the measurable quantity is the probability that a foul is called. This rate of fouls is:

$$\Pr[\hat{h}(Y) = h_1] = 1 - p_0 + p_0 p_E^{\mathrm{I}} - (1 - p_0) p_E^{\mathrm{II}}. \quad (7)$$

Looking at the average performance of a referee over the populations of black and white players, we compare the expected foul rates on whites and blacks:

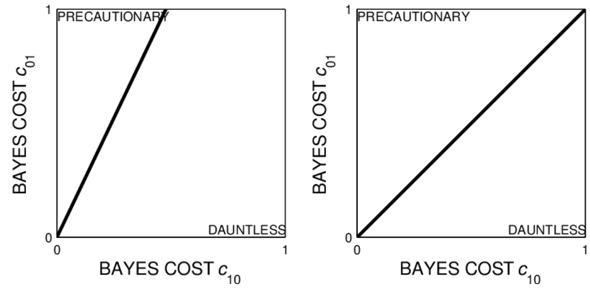
$$\begin{aligned}\Delta(c_{10}, c_{01}) &= E\left[\Pr[\hat{h}_{K_b}(Y) = h_1] - \Pr[\hat{h}_{K_w}(Y) = h_1]\right] \\ &= E[P_0 p_E^{\mathrm{I}}(q^*_{K_b}(P_0)) - (1 - P_0) p_E^{\mathrm{II}}(q^*_{K_b}(P_0)) \\ &\quad - P_0 p_E^{\mathrm{I}}(q^*_{K_w}(P_0)) + (1 - P_0) p_E^{\mathrm{II}}(q^*_{K_w}(P_0))]. \quad (8)\end{aligned}$$

If this discrimination quantity Δ is greater than zero, then the referee is calling more fouls on blacks. If Δ is less than zero, then the referee is calling more fouls on whites.

The dependence of Δ on c_{10} and c_{01} has been explicitly notated on the left side of (8) and is implicit in the two types of error probabilities on the right side of (8). The value of Δ also depends on the unquantized prior distribution $f_{P_0}(p_0)$, the values of K_w and K_b, and the measurement model. Fixing these, we can determine the regions in the c_{10}–c_{01} plane where a referee would call more fouls on blacks and where a referee would call more fouls on whites. This is shown in Fig. 7. For the uniform prior $f_{P_0}(p_0)$, the two regions are divided by the line $c_{01} = c_{10}$. For the beta distribution prior depicted, the dividing line is $c_{01} = \alpha c_{10}$, where $\alpha > 1$.

For any population model and measurement model, there is one half-plane where a referee would call more fouls on black players and the other half-plane where the referee would call more fouls on white players. To reiterate, just because the Bayes risk for foul-calling on black players is greater than that for white players, it does not automatically imply that the foul call rate for blacks is higher. The high Bayes risk could well be the result of a preponderance of missed foul calls.

This result may be interpreted in terms of precautionary and dauntless decision making. The precautionary principle corresponds to Bayes cost assignment with $c_{01} \gg c_{10}$, whereas the dauntless principle corresponds to Bayes cost assignment with $c_{01} \ll c_{10}$. Thus, we may call a referee with $K_w > K_b$ that calls more fouls on black players as precautionary and that calls more fouls on white players as dauntless. A referee with $K_w < K_b$ that calls more fouls on black players is dauntless and more fouls on white players is precautionary.



(a) (b)

Fig. 7. Dividing line between Bayes cost region in which referee will call more fouls on blacks and region in which referee will call more fouls on whites. A referee with $K_b < K_w$ will call more fouls on blacks in the upper left region and more fouls on whites in the lower right region, which correspond to precautionary and dauntless respectively. For (a), the prior probability distribution is beta(5,2), measurements are through additive white Gaussian noise with unit signal to noise ratio, and level allocation is $K_b = 3$, $K_w = 4$. For (b), the prior probability distribution is uniform.

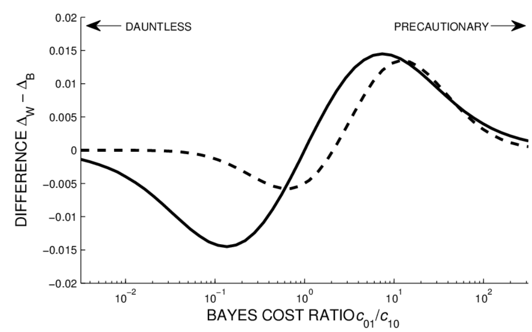


Fig. 8. Difference in differences in foul calling as a function of the Bayes cost ratio. The white referee has $K_w = 4$, $K_b = 3$ and the black referee has $K_w = 3$, $K_b = 4$. For the dashed line, the prior probability distribution is beta(5,2) and measurements are through additive white Gaussian noise with unit signal to noise ratio. For the solid line, the prior probability distribution is uniform.

Econometric studies often give differences in differences to show racial bias. The first 'difference' is the difference in foul call rate between black players and white players, which in our notation is Δ. The second 'difference' is the difference in Δ between white referees and black referees. Denoting the foul call rate difference of a white referee by Δ_W and the foul call rate difference of a black referee by Δ_B, the difference in differences is $\Delta_W - \Delta_B$.

Fig. 8 plots the difference in differences as a function of the ratio c_{01}/c_{10} for two different population distributions, a beta distribution and the uniform distribution. The right side of the plot is the precautionary regime, where white referees would call more fouls on black players than black referees. For the particular examples, if $c_{01}/c_{10} = 10$, then the white referee has a foul call rate 0.0132 greater than the black referee on black players for the beta distribution and 0.0142 greater for the uniform distribution.

The left side of the plot is the dauntless regime, where white referees would call fewer fouls on black players than black referees. For the particular examples, if $c_{01}/c_{10} = 0.1$, then the white referee has a foul call rate 0.0013 less than the black referee on black players for the beta distribution and 0.0142 less for the uniform distribution. In these examples, the white referee has $K_w = 4$, $K_b = 3$, and the black referee has $K_w = 3$, $K_b = 4$.[4]

C. Price of Segregation

Decision makers of different races exhibit different biases because they have different K_w and K_b allocations due to different $m_w/(m_w + m_b)$ ratios. This ratio is not the actual fraction of whites whose actions are assessed by the decision maker, but is determined in part by the decision maker's segregated social life. If decision makers of all races have a bias that matches the true white fraction, then the phenomenon of racial bias would actually achieve optimal social welfare.[5] Different decision-making biases by different decision makers, however, cannot simultaneously be societally optimal.

Our model fixes limitations of human information processing, automaticity of racial classification, and intertwining of social and economic life. Social segregation causes mismatch between social and economic lives and is therefore the root cause of non-optimal racial bias. To draw connections to econometric studies where ground truth is not known, the previous section used differences in differences. In analogy with notions of welfare loss in economic theory like deadweight loss, the social cost of monopoly [88, Ch. 4], and the price of anarchy [89], a *price of segregation* is defined here as a way to measure the deleterious effect of segregation.

Let $\pi_{\rm true}$ be the fraction of whites in the economic decision-making setting. A particular decision maker that leads a segregated life, on the other hand, will have a white ratio $\pi_{\rm seg} = m_w/(m_w + m_b)$. The mean BRE, from the perspective of society, under the true white fraction is

$$D^{(2)}(\pi_{\rm true}) = \pi_{\rm true} E[J(P_0, q_{K_w(\pi_{\rm true})}(P_0))] \\ + (1-\pi_{\rm true}) E[J(P_0, q_{K_b(\pi_{\rm true})}(P_0))] - E[J(P_0, P_0)]$$

whereas the mean BRE, from the perspective of society, under the segregated white fraction is

$$D^{(2)}(\pi_{\rm seg}) = \pi_{\rm true} E[J(P_0, q_{K_w(\pi_{\rm seg})}(P_0))] \\ + (1-\pi_{\rm true}) E[J(P_0, q_{K_b(\pi_{\rm seg})}(P_0))] - E[J(P_0, P_0)].$$

The difference between these two is the price of segregation:

$$\Pi = D^{(2)}(\pi_{\rm true}) - D^{(2)}(\pi_{\rm seg}) \\ = \pi_{\rm true} \left\{ E[J(P_0, q_{K_w(\pi_{\rm true})}(P_0))] - E[J(P_0, q_{K_w(\pi_{\rm seg})}(P_0))] \right\} \\ + (1-\pi_{\rm true}) \left\{ E[J(P_0, q_{K_t - K_w(\pi_{\rm true})}(P_0))] \right. \\ \left. - E[J(P_0, q_{K_t - K_w(\pi_{\rm seg})}(P_0))] \right\}.$$

[4]There is no requirement for the white referee to have $K_w > K_b$ and the black referee to have $K_w < K_b$. It is only required that the K_w of the white referee be greater than the K_w of the black referee (assuming the same K_t). We get a plot qualitatively similar to Fig. 8 if for example the white referee has $K_w = 5$, $K_b = 2$, and the black referee has $K_w = 4$, $K_b = 3$.

[5]This phenomenon is precisely statistical discrimination; its social welfare optimality is why courts have ruled it legal.

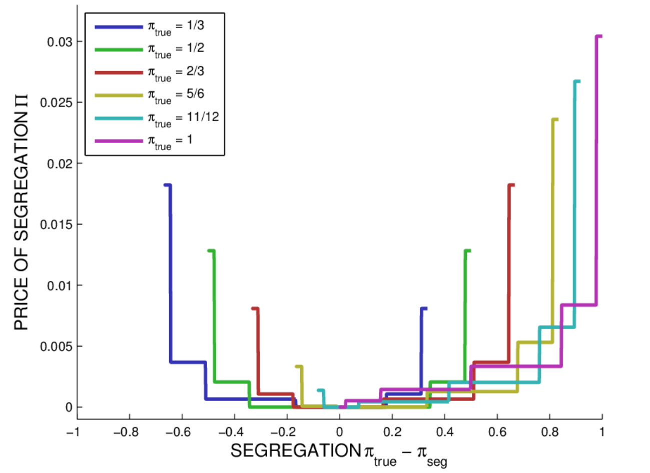


Fig. 9. The price of segregation Π as a function of the level of segregation mismatch $\pi_{\rm true} - \pi_{\rm seg}$ for several values of $\pi_{\rm true}$. The distribution of the prior probability is beta(5,2), the measurements are observed through additive white Gaussian noise with unit signal to noise ratio, the Bayes costs are $c_{10} = c_{01} = 1$, and $K_t = 7$.

The price of segregation Π depends strongly on the discontinuous, integer-valued $K_w(\cdot)$ function, and is also discontinuous. The price of segregation is a non-decreasing function of the level of segregation mismatch $|\pi_{\rm true} - \pi_{\rm seg}|$. An example of the price of segregation for a particular society and several values of $\pi_{\rm true}$ is shown in Fig. 9. Notice that if the level of mismatch is small, there may be no price of segregation.

The model predicts that greater homogeneity of social interaction among people would mitigate the price of segregation by driving the $\pi_{\rm seg}$ for all decision makers closer to $\pi_{\rm true}$.

VI. Conclusion

In this paper, we have formulated a model of LRT detection with quantized priors which generates discriminative behavior when combined with theories of social cognition and facts about social segregation. The discriminative behavior arises despite having identical distributions for different populations and despite no malicious intent on the part of the decision maker. The quantization of priors model is developed not only because "categorizing serves to cut down the diversity of objects and events that must be dealt with uniquely by an organism of limited capacities" [90, p. 235], but also because limiting complexity promotes good generalization and prevents overfitting in learning [30, Ch. 5.3].

A. Empirical Comparisons

The proposed model predicts that decision-making performance of a decision maker of the opposite race as the citizen or referee is worse than performance of a decision maker of the same race. This prediction is in fact born out experimentally. A large literature in face recognition shows exactly the own-race bias effect that we predict, observed colloquially as "they [other-race persons] all look alike." In particular, both kinds of errors that factor into Bayes risk increase when trying to recognize members of the opposite population [91], as verified with laboratory experiments.

Econometric studies also provide a source for comparison to the proposed model, but as noted, only $\Pr[\hat{h}(Y) = h_1]$ is observed. We note that the addition of police officers of a given race is associated with an increase in the number of arrests of suspects of a different race but has little impact on same-race arrests. The effect is more pronounced for minor offenses [20] where the prior probability presumably plays a bigger role than the measurement. There are similar own-race bias effects in the decision by police to search a vehicle during a traffic stop [21] and in the decision of NBA referees to call a foul [19]. The rate of searching and the rate of foul calling are greater when the decision maker is of a different race than the driver and player, respectively. These studies are consistent with model predictions if decision makers are precautionary.

For example, [21] report $\Pr[\hat{h}(Y) = h_1] = 0.0097$ for white officers and black drivers, $\Pr[\hat{h}(Y) = h_1] = 0.0040$ for white officers and white drivers, $\Pr[\hat{h}(Y) = h_1] = 0.0082$ for black officers and black drivers, and $\Pr[\hat{h}(Y) = h_1] = 0.0062$ for black officers and white drivers. Thus, $\Delta_W = 0.0097 - 0.0040 = 0.0057$ and $\Delta_B = 0.0082 - 0.0062 = 0.0020$, resulting in a difference in differences of 0.0037. This is of the same order of magnitude seen in Fig. 8 with precautionary cost settings. Interestingly, there are two different precautionary cost settings that yield a given difference in differences value: one in which the cost ratio is fairly small and another in which the cost ratio is quite large. An exact numerical prediction for the difference in differences curve could be made from the model if we had knowledge of $f_{P_0}(p_0)$, K_w, and K_b.

An experiment found jurors in a simulated rape trial in Canada convicted same-race defendants more than other-race defendants, though there were confounding variables [46]. This study is consistent with model predictions if decision makers are dauntless. A precautionary utility function leads to higher $\hat{h} = h_1$ rates for the opposite race whereas a dauntless view leads to higher rates for the own race. Such a phenomenon of the decision maker's attitude fundamentally altering the nature of discrimination seems not to have been described before. The Bayes costs of decision makers are revealed in their bias; the model suggests Canadian jurors truly do believe in the standard of proof and the concept of "innocent until proven guilty."

Note that utility elicitation may be used to determine whether a decision maker is precautionary or dauntless [92]. As such, our theoretical framework is empirically falsifiable by observed inconsistency between attitude and bias.

B. Testing for Quantization-Based Discrimination

Due to legal and policy considerations, economists have developed several empirical tests to differentiate statistical discrimination (in the strict sense of [33]) from taste-based discrimination [21], [63], [64]. These tests are designed to determine whether there are differences between racial populations in the dimension relevant to the decision-making task or not. If the test finds no basis for statistical discrimination, then the discrimination is attributed to taste-based discrimination.

Our model assumes there is no difference between racial populations and therefore no basis for statistical discrimination. Our information-based discrimination can be differentiated from statistical discrimination using the same tests as for taste-based discrimination. Indeed, our model of decision making provides an alternate explanation for racial bias that has traditionally been attributed to a taste for discrimination.

Both our information-based discrimination and taste-based discrimination can be differentiated from statistical discrimination on the basis of econometric data, but one might wonder whether the two can be differentiated from each other. Recall the expression for the threshold in the LRT (1): $c_{10}a/c_{01}(1 - a)$. Taste-based discrimination implies different Bayes costs c_{ij} for citizens of different races whereas information-based discrimination may imply different prior belief parameters a for citizens of different races. Since Bayes costs and prior belief parameter are mixed together in the threshold, their individual effects cannot be discerned without independent data about c_{10}/c_{01} or about $a/(1 - a)$. As Radner remarked [93], "differences of opinion (as embodied in different a priori distributions, for example)…[are equivalent to] conflicts of interest," i.e. a taste for discrimination.

The courts have often declared taste-based discrimination illegal but have found rational statistical discrimination legal [21], [63], [94], since statistical discrimination is not the product of animus [95]. Here we have developed a model of information-based discrimination free from animus, that is often empirically indistinguishable from taste-based discrimination. Hence the legal evidentiary standard for taste-based discrimination may need to be revisited.

As a final note, implicit discrimination [34] is said to arise due to implicit pro-white/anti-black cognitive associations among *both* white and black decision makers [96]–[99]. It is very different from our information-based discrimination and from taste-based discrimination. Due to its asymmetry, it should be readily differentiable in empirical data.

C. Model Extensions

The present model of decision making can be extended to higher dimensions of various types. First, one may consider M-ary hypothesis testing rather than just binary, e.g. for a stage in sequential hypothesis testing, or in college admissions where outcomes include $\{$*admit*, *reject*, *waitlist*$\}$. The resultant vector quantization problem partitions the $M - 1$ dimensional probability simplex. Second, one may increase the number of racial groups from two to N. An exhaustive search over representation point allocations for integer partitions of K_t: $K_t = K_1 + K_2 + \cdots + K_N$ may be used; this only involves the design of a linear number of quantizers and a small optimization problem. Third, one may consider effects of a higher-dimensional socially-observable attribute space; identity is not just race. In fact some social dimensions may be consciously and explicitly correlated in order to further define identity [77]. The gains of vector quantization over scalar quantization are enhanced when there is dependence among dimensions [4].

Besides extensions to higher dimensions, one can consider a restricted class of quantizers rather than optimal quantization.

Such restriction may model further cognitive constraints on decision makers. In particular, Fryer and Jackson [28] have suggested a heuristic algorithm for quantizer design based on splitting groups, which is a rediscovery of the tree-structured vector quantizer design algorithm given by [100, Fig. 20].

One may also consider group decision making, where team-theoretic and game-theoretic considerations arise [11], [101].

D. Discussion

Discrimination on the basis of race has been a troublesome problem in civic life. This paper has presented a formal model of decision making reflecting ideas from psychology, social cognition theory, and sociology, which generates such discriminative behavior. It includes full Bayesian rationality as a limiting case that displays no discrimination. Biased decision making arises despite having identical *ex ante* propensity distributions for different populations and despite no taste for discrimination or implicit discrimination on the part of the decision maker.

In a sense, the model predicts that greater homogeneity of social interaction among decision makers would mitigate the price of segregation. This draws a connection to one branch of intergroup contact theory, which suggests that contact reduces prejudice since it allows individuals the chance to see previously unnoticed similarities and counter-stereotypic characteristics and behaviors in one another [102], [103], a conclusion similar to model predictions. Perhaps unexpectedly, social interaction is not linear in the overall ratio of subgroup populations [42], perhaps due to structural reasons [35].

We have drawn on the science of information to suggest a mechanism by which discrimination may arise despite lack of malevolence and *a priori* identicality between populations. Discrimination in the model arises through assumptions of automaticity of classification along social characteristics and the finite human capacity for information processing. This complements taste-based, statistical, and implicit mechanisms that explain racial discrimination.

Acknowledgements

Thanks to V. K. Goyal, S. K. Mitter, A. S. Willsky, D. Acemoglu, and S. Mullainathan for discussions.

References

[1] W. Cronon, *Nature's Metropolis: Chicago and the Great West*. New York: W. W. Norton & Company, 1991.

[2] G. Pagès, H. Pham, and J. Printems, "Optimal quantization methods and applications to numerical problems in finance," in *Handbook of Computational and Numerical Methods in Finance*, S. T. Rachev, Ed. Boston: Birkhäuser, 2004, pp. 253–297.

[3] G. S. Becker, *A Treatise on the Family*, enlarged ed. Cambridge, MA: Harvard University Press, 1991.

[4] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

[5] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[6] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2693–2723, Oct. 1998.

[7] C. A. Sims, "Implications of rational inattention," *J. Monet. Econ.*, vol. 50, no. 3, pp. 665–690, Apr. 2003.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July/Oct. 1948.

[9] K. R. Varshney and L. R. Varshney, "Quantization of prior probabilities for hypothesis testing," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4553–4562, Oct. 2008.

[10] ——, "Optimal grouping for group minimax hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6511–6521, Oct. 2014.

[11] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Quantization of prior probabilities for collaborative distributed hypothesis testing," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4537–4550, Sep. 2012.

[12] ——, "Distributed decision making by categorically-thinking agents," in *Decision Making and Imperfection*, T. V. Guy, M. Karny, and D. H. Wolpert, Eds. Springer, 2013, pp. 37–63.

[13] B. M. Turner, T. Van Zandt, and S. Brown, "A dynamic stimulus-driven model of signal detection," *Psychol. Rev.*, vol. 118, no. 4, pp. 583–613, Oct. 2011.

[14] F. P. Ramsey, "Truth and probability," in *The Foundations of Mathematics and Other Logical Essays*, R. B. Braithwaite, Ed. New York: Harcourt, Brace and Company, 1931, pp. 156–198.

[15] B. de Finetti, "La prévision: Ses lois logiques, ses sources subjectives," *Ann. Inst. Henri Poincaré*, vol. 7, no. 1, pp. 1–68, 1937.

[16] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press, 1944.

[17] L. J. Savage, *The Foundations of Statistics*. New York: Wiley, 1954.

[18] A. S. Willsky, G. W. Wornell, and J. H. Shapiro, *Stochastic Processes, Detection and Estimation 6.432 Course Notes*. Cambridge, MA: Dept. Elect. Eng. Comput. Sci., Mass. Inst. Tech., Fall 2003.

[19] J. Price and J. Wolfers, "Racial discrimination among NBA referees," *Quart. J. Econ.*, vol. 125, no. 4, pp. 1859–1887, Nov. 2010.

[20] J. J. Donohue, III and S. D. Levitt, "The impact of race on policing and arrests," *J. Law Econ.*, vol. 44, no. 2, pp. 367–394, Oct. 2001.

[21] K. L. Antonovics and B. G. Knight, "A new look at racial profiling: Evidence from the Boston Police Department," *Rev. Econ. Stat.*, vol. 91, no. 1, pp. 163–177, Feb. 2009.

[22] J. A. Swets, W. P. Tanner, Jr., and T. G. Birdsall, "Decision processes in perception," *Psychol. Rev.*, vol. 68, no. 5, pp. 301–340, Sep. 1961.

[23] M. Glanzer, A. Hilford, and L. T. Maloney, "Likelihood ratio decisions in memory: Three implied regularities," *Psychon. Bull. Rev.*, vol. 16, no. 3, pp. 431–455, Jun. 2009.

[24] G. L. Brase, L. Cosmides, and J. Tooby, "Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty," *J. Exp. Psychol. Gen.*, vol. 127, no. 1, pp. 3–21, Mar. 1998.

[25] W. K. Viscusi, "Are individuals Bayesian decision makers?" *Am. Econ. Rev.*, vol. 75, no. 2, pp. 381–385, May 1985.

[26] T. L. Griffiths, C. Kemp, and J. B. Tenenbaum, "Bayesian models of cognition," in *The Cambridge Handbook of Computational Cognitive Modeling*, R. Sun, Ed. New York: Cambridge University Press, 2008, pp. 59–100.

[27] G. A. Miller, "Human memory and the storage of information," *IRE Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 129–137, Sep. 1956.

[28] R. Fryer and M. O. Jackson, "A categorical model of cognition and biased decision-making," *B. E. J. Theor. Econ.*, vol. 8, no. 1, Jan. 2008.

[29] S. Mullainathan, J. Schwartzstein, and A. Shleifer, "Coarse thinking and persuasion," *Quart. J. Econ.*, vol. 123, no. 2, pp. 577–619, May 2008.

[30] K. R. Varshney, "Frugal hypothesis testing and classification," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, Jun. 2010.

[31] I. Gilboa, *Theory of Decision Under Uncertainty*. Cambridge: Cambridge University Press, 2009.

[32] G. S. Becker, *The Economics of Discrimination*. Chicago: University of Chicago Press, 1957.

[33] K. Arrow, "The theory of discrimination," in *Discrimination in Labor Markets*, O. Ashenfelter and A. Rees, Eds. Princeton, NJ: Princeton University Press, 1973, pp. 3–33.

[34] M. Bertrand, D. Chugh, and S. Mullainathan, "Implicit discrimination," *Am. Econ. Rev.*, vol. 95, no. 2, pp. 94–98, May 2005.

[35] P. Blau, *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press, 1977.

[36] M. Augoustinos, I. Walker, and N. Donaghue, *Social Cognition: An Integrated Introduction*, 2nd ed. London: SAGE Publications, 2006.

[37] C. N. Macrae and G. V. Bodenhausen, "Social cognition: Thinking categorically about others," *Annu. Rev. Psychol.*, vol. 51, pp. 93–120, Feb. 2000.

[38] L. Cosmides, J. Tooby, and R. Kurzban, "Perceptions of race," *Trends Cogn. Sci.*, vol. 7, no. 4, pp. 173–179, Apr. 2003.

[39] D. Chugh, “Societal and managerial implications of implicit social cognition: Why milliseconds matter,” *Social Justice Research*, vol. 17, no. 2, pp. 203–222, Jun. 2004.

[40] J. N. Wood, “Social cognition and the prefrontal cortex,” *Behav. Cogn. Neurosci. Rev.*, vol. 2, no. 2, pp. 97–114, Jun. 2003.

[41] D. S. Massey and N. A. Denton, “The dimensions of residential segregation,” *Soc. Forces*, vol. 67, no. 2, pp. 281–315, Dec. 1988.

[42] F. Echenique and R. G. Fryer, Jr., “A measure of segregation based on social interactions,” *Quart. J. Econ.*, vol. 122, no. 2, pp. 441–485, May 2007.

[43] R. D. Putnam, “E pluribus unum: Diversity and community in the twenty-first century,” *Scand. Polit. Stud.*, vol. 30, no. 2, pp. 137–174, Jun. 2007.

[44] M. C. Taylor, “How white attitudes vary with the racial composition of local populations: Numbers count,” *Am. Sociol. Rev.*, vol. 63, no. 4, pp. 512–535, Aug. 1998.

[45] L. Quillian, “Prejudice as a response to perceived group threat: Population composition and anti-immigrant and racial prejudice in Europe,” *Am. Sociol. Rev.*, vol. 60, no. 4, pp. 586–611, Aug. 1995.

[46] R. M. Bagby, J. D. Parker, N. A. Rector, and V. Kalemba, “Racial prejudice in the Canadian legal system: Juror decisions in a simulated rape trial,” *L. Hum. Behav.*, vol. 18, no. 3, pp. 339–350, Jun. 1994.

[47] J. Conlisk, “Why bounded rationality?” *J. Econ. Lit.*, vol. 34, no. 2, pp. 669–700, Jun. 1996.

[48] A. Rubinstein, *Modeling Bounded Rationality*. Cambridge, MA: MIT Press, 1998.

[49] R. Radner, “Costly and bounded rationality in individual and team decision-making,” *Ind. Corporate Change*, vol. 9, no. 4, pp. 623–658, Dec. 2000.

[50] A. d. Monteiro, “Similarity and categorization: A rational strategy to simplify problems,” Jan. 2007.

[51] S. Mullainathan, “A memory-based model of bounded rationality,” *Quart. J. Econ.*, vol. 117, no. 3, pp. 735–774, Aug. 2002.

[52] Y. Sarafidis, “What have you done for me lately? release of information and strategic manipulation of memories,” *Econ. J. (London)*, vol. 117, no. 518, pp. 307–326, Mar. 2007.

[53] J. Dow, “Search decisions with limited memory,” *Rev. Econ. Stud.*, vol. 58, no. 1, pp. 1–14, Jan. 1991.

[54] M. Piccione and A. Rubinstein, “On the interpretation of decision problems with imperfect recall,” *Games Econ. Behav.*, vol. 20, no. 1, pp. 3–24, Jul. 1997.

[55] A. Wilson, “Bounded memory and biases in information processing,” Apr. 2003.

[56] D. Monte, “Reputation and bounded memory: A theory of inertia and skepticism,” Dec. 2007.

[57] S. Mullainathan, “Thinking through categories,” Apr. 2002.

[58] J. E. Stiglitz, “Information and the change in the paradigm in economics,” Nobel Prize Lecture, Dec. 2001.

[59] E. S. Phelps, “The statistical theory of racism and sexism,” *Am. Econ. Rev.*, vol. 62, no. 4, pp. 659–661, Sep. 1972.

[60] M. Peski, “Categorization,” Jan. 2005.

[61] M. Spence, “Job market signaling,” *Quart. J. Econ.*, vol. 87, no. 3, pp. 355–374, Aug. 1973.

[62] S. Coate and G. C. Loury, “Will affirmative-action policies eliminate negative stereotypes?” *Am. Econ. Rev.*, vol. 83, no. 5, pp. 1220–1240, Dec. 1993.

[63] J. Knowles, N. Persico, and P. Todd, “Racial bias in motor vehicle searches: Theory and evidence,” *J. Polit. Econ.*, vol. 109, no. 1, pp. 203–229, Feb. 2001.

[64] S. Anwar and H. Fang, “An alternative test of racial prejudice in motor vehicle searches: Theory and evidence,” *Am. Econ. Rev.*, vol. 96, no. 1, pp. 127–151, Mar. 2006.

[65] D. J. Aigner and G. G. Cain, “Statistical theories of discrimination in labor markets,” *Ind. Lab. Relat. Rev.*, vol. 30, no. 2, pp. 175–187, Jan. 1977.

[66] S. J. Lundberg and R. Startz, “Private discrimination and social intervention in competitive labor market,” *Am. Econ. Rev.*, vol. 73, no. 3, pp. 340–347, Jun. 1983.

[67] B. Cornell and I. Welch, “Culture, information, and screening discrimination,” *J. Polit. Econ.*, vol. 104, no. 3, pp. 542–571, Jun. 1996.

[68] E. Rosch, “Principles of categorization,” in *Cognition and Categorization*, E. Rosch and B. B. Lloyd, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1978, pp. 27–48.

[69] R. L. Goldstone, “The role of similarity in categorization: Providing a groundwork,” *Cognition*, vol. 52, no. 2, pp. 125–157, Aug. 1994.

[70] K. R. Varshney, “Bayes risk error is a Bregman divergence,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4470–4472, Sep. 2011.

[71] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

[72] H. Steinhaus, “Sur la division des corp materiels en parties,” *Bull. Acad. Polonaise Sci.*, vol. C1.III, no. IV, pp. 801–804, 1956.

[73] S. P. Lloyd, “Least squares quantization in PCM,” Jul. 1957, unpublished Bell Laboratories Technical Note.

[74] J. Max, “Quantizing for minimum distortion,” *IRE Trans. Inf. Theory*, vol. IT-6, no. 1, pp. 7–12, Mar. 1960.

[75] A. Banerjee, X. Guo, and H. Wang, “On the optimality of conditional expectation as a Bregman predictor,” *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2664–2669, Jul. 2005.

[76] A. V. Trushkin, “Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions,” *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 187–198, Mar. 1982.

[77] G. A. Akerlof and R. E. Kranton, “Economics and identity,” *Quart. J. Econ.*, vol. 115, no. 3, pp. 715–753, Aug. 2000.

[78] Y. Shoham and A. Gersho, “Efficient bit allocation for an arbitrary set of quantizers,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 9, pp. 1445–1453, Sep. 1988.

[79] V. N. Vapnik, *The Nature of Statistical Learning*. New York: Springer-Verlag, 2000.

[80] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Oct. 2005.

[81] R. M. Gray, J. C. Kieffer, and Y. Linde, “Locally optimal block quantizer design,” *Inf. Control*, vol. 45, no. 2, pp. 178–198, May 1980.

[82] M. J. Sabin and R. M. Gray, “Global convergence and empirical consistency of the generalized Lloyd algorithm,” *IEEE Trans. Inf. Theory*, vol. IT-32, no. 2, pp. 148–155, Mar. 1986.

[83] M. C. du Plessis and M. Sugiyama, “Semi-supervised learning of class balance under class-prior change by distribution matching,” in *Proc. 29th Int. Conf. Mach. Learn. (ICML 2012)*, Jun.–Jul. 2012, pp. 823–830.

[84] T. Sanderson and C. Scott, “Class proportion estimation with application to multiclass anomaly rejection,” in *Proc. 17th Int. Conf. Artif. Intell. Stat. (AISTATS 2014)*, Apr. 2014, pp. 850–858.

[85] C. Scott, “A rate of convergence for mixture proportion estimation, with application to learning from noisy labels,” in *Proc. 18th Int. Conf. Artif. Intell. Stat. (AISTATS 2015)*, May 2015, pp. 838–846.

[86] M. C. du Plessis, G. Niu, and M. Sugiyama, “Class-prior estimation for learning from positive and unlabeled data,” in *Proc. 7th Asian Conf. Mach. Learn. (ACML 2015)*, Nov. 2015, pp. 221–236.

[87] C. Brunet and S. Loustau, “Noisy quantization: Theory and practice,” hal-01060380 [math.ST], Jul. 2014.

[88] W. K. Viscusi, J. E. Jr., Harrington, and J. M. Vernon, *Economics of Regulation and Antitrust*. Cambridge, MA: MIT Press, 2005.

[89] C. H. Papadimitriou, “Algorithms, games, and the internet,” in *Proc. 33rd Annu. ACM Symp. Theory Comput. (STOC’01)*, Jul. 2001, pp. 749–753.

[90] J. S. Bruner, J. J. Goodnow, and G. A. Austin, *A Study of Thinking*. New York: John Wiley & Sons, 1956.

[91] C. A. Meissner and J. C. Brigham, “Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review,” *Psychol. Pub. Pol. L.*, vol. 7, no. 1, pp. 3–35, Jan. 2001.

[92] P. Wakker and D. Deneffe, “Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown,” *Manage. Sci.*, vol. 42, no. 8, pp. 1131–1150, Aug. 1996.

[93] R. Radner, “Team decision problems,” *Ann. Math. Stat.*, vol. 33, no. 3, pp. 857–881, Sep. 1962.

[94] A. Farmer and D. Terrell, “Crime versus justice: Is there a trade-off?” *J. Law Econ.*, vol. 44, no. 2, pp. 345–366, Oct. 2001.

[95] D. A. Strauss, ““group rights” and the problem of statistical discrimination,” *Issues in Legal Scholarship: The Origins and Fate of Antisubordination Theory*, p. 17, 2003.

[96] E. A. Phelps, K. J. O’Connor, W. A. Cunningham, E. S. Funayama, J. C. Gatenby, J. C. Gore, and M. R. Banaji, “Performance on indirect measures of race evaluation predicts amygdala activation,” *J. Cogn. Neurosci.*, vol. 12, no. 5, pp. 729–738, Sep. 2000.

[97] A. S. Baron and M. R. Banaji, “The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood,” *Psychol. Sci.*, vol. 17, no. 1, pp. 53–58, Jan. 2006.

[98] B. A. Nosek, M. R. Banaji, and A. G. Greenwald, “Harvesting implicit group attitudes and beliefs from a demonstration web site,” *Group Dyn. Theor. Research Practice*, vol. 6, no. 1, pp. 101–115, Mar. 2002.

[99] J. T. Jost, M. R. Banaji, and B. A. Nosek, “A decade of system justification theory: Accumulated evidence of conscious and unconscious

bolstering of the status quo," *Polit. Psychol.*, vol. 25, no. 6, pp. 881–919, Dec. 2004.

[100] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, no. 11, pp. 1551–1588, Nov. 1985.

[101] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Conflict in distributed hypothesis testing with quantized prior probabilities," in *Proc. IEEE Data Compression Conf. (DCC 2011)*, Mar. 2011, pp. 313–322.

[102] M. Rothbart and O. P. John, "Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact," *J. Soc. Issues*, vol. 41, no. 3, pp. 81–104, Fall 1985.

[103] C. Van Laar, S. Levin, S. Sinclair, and J. Sidanius, "The effect of university roommate contact on ethnic attitudes and behavior," *J. Exp. Soc. Psychol.*, vol. 41, no. 4, pp. 329–345, Jul. 2005.

Lav R. Varshney (S'00–M'10–SM'15) received the B.S. degree (*magna cum laude*) in electrical and computer engineering with honors from Cornell University, Ithaca, New York, in 2004. He received the S.M., E.E., and Ph.D. degrees, all in electrical engineering and computer science, from the Massachusetts Institute of Technology, Cambridge, in 2006, 2008, and 2010, where his theses received the E. A. Guillemin Thesis Award and the J.-A. Kong Award Honorable Mention.

He is an assistant professor in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Beckman Institute at the University of Illinois at Urbana-Champaign. During 2010–2013, he was a research staff member at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York. His research interests include information and coding theory; limits of nanoscale, human, and neural computing; human decision making and collective intelligence; and creativity.

Dr. Varshney is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He and his students have won several best paper awards including the 2013 Best Student Paper Award of the IEEE Data Storage Technical Committee and the 2015 Data for Good Exchange Paper Award. His work appears in the anthology, *The Best Writing on Mathematics 2014* (Princeton University Press). He is a founding member of the IEEE Special Interest Group on Big Data in Signal Processing and currently serves on the Shannon Centenary Committee of the IEEE Information Theory Society.

Kush R. Varshney (S'00–M'10–SM'15) received the B.S. degree (magna cum laude) in electrical and computer engineering with honors from Cornell University, Ithaca, New York, in 2004. He received the S.M. degree in 2006 and the Ph.D. degree in 2010, both in electrical engineering and computer science at the Massachusetts Institute of Technology (MIT), Cambridge.

He is a research staff member in the Data Science Group of the Mathematical Sciences Department at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York and codirector of the IBM Social Good Fellowship program. His research interests include statistical signal processing, machine learning, data mining, and image processing.

Dr. Varshney is a member of Eta Kappa Nu and Tau Beta Pi. He received a Best Student Paper Travel Award at the 2009 International Conference on Information Fusion, the Best Paper Award at the 2013 IEEE International Conference on Service Operations and Logistics, and Informatics, the Best Social Good Paper Award at the 2014 ACM SIGKDD Conference on Knowledge Discovery and Data Mining, a Best Research Paper Honorable Mention at the 2015 SIAM International Conference on Data Mining, and several IBM awards for contributions to analytics projects. He is a member of the IEEE Signal Processing Society's Machine Learning for Signal Processing Technical Committee and Signal Processing Theory and Methods Technical Committee.