Bias in the State: Theory and Evidence from Peru*

Michael C. Best
Columbia U.
& BREAD & CEPR & NBER

Leonard Goff
U. Calgary

Jonas Hjort
UCL
& BREAD & CEPR & U. Oslo

Dafne Murillo
Columbia U.

Gastón Pierri
Inter American Development Bank

This version: March 26, 2026

Abstract

We study how beliefs and preferences in states' organizations generate differential treatment of minorities and what policy can do about it. Our framework decomposes each agent's choices into (accurate) statistical discrimination, incorrect-beliefs bias, and taste-based discrimination. The framework is overidentified and thus allows a test of whether behavior can be rationalized. Empirically, "sequential benchmarking"—actions against beliefs, beliefs against actual downstream outcomes—identifies the various elements of agents' choices. To illustrate, we apply the framework to the agency handling corruption accusations in Peru: its most salient, high-stakes citizen–state interactions. The *Contraloría*'s mission (indictments) is defined by law and agents' information set (whistleblower reports) observable. The framework maps to a particular experimental design. Agents triage known-(to us)-outcome case files and report (0–100) both their perceived probability of indictment given escalation and priority for escalation. Varying the ethnicity—European vs. indigenous—on some cases gives us a legally defined "ethnicity itself" measure of the trait and agents' responses capture the decomposition's starting point: their *overall* differential treatment. Most, but not all, display rationalizable but discriminatory behavior. They overestimate indictment risk and do so more for indigenous defendants; belief errors attenuate measured statistical discrimination and, so (counterfactually) assuming correct beliefs would flip the sign and magnitude of estimated taste-based bias. Bias varies substantially across agents, implying distinct remediation (e.g. assignment) policies and outcomes for principals who seek to maximize indictments subject to minimal taste-based bias; taste-based and incorrect-beliefs bias; or any form of differential treatment. Our framework and estimates outline how to reduce differential treatment while preserving states' goals.

*Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the Inter-American Development Bank or La Contraloría General de la República del Perú. Neither IADB nor CGR can be held responsible for them. We thank the National Science Foundation, the International Growth Centre, and the Stone Centre at UCL for funding; Mirko Daga Acevedo, Marco Gutiérrez, Alexandra Marcos Quispe, Marcia Ruiz, Sebastián Sardón, and Bennett Smith-Worthington for excellent research assistance; and audience members at many seminars and conferences for great feedback.

1 Introduction

Research in economics often attempts to separate the various dimensions of agents' choices—their overall goals, elements of their preferences, beliefs, and so on. This is typically necessary to model decision-making and to draw policy implications from observed behavior. An especially consequential example is discrimination against ethnic minorities by state entities. Uniform treatment of citizens is enshrined in most constitutions, and the potential economic cost and individual welfare loss from differential treatment are large (Arnold et al., 2018; Hsieh et al., 2019; Goncalves and Mello, 2021). Not all forms of seemingly differential treatment reflect *bias* in decision-makers' goals, preferences, or beliefs, however. Even if both the targeted overall outcome and the constraints state officials operate under are known, many different combinations of (potentially skewed) preferences over subpopulations' wellbeing and (potentially incorrect) beliefs about subpopulations can often rationalize a given behavior. In other words, "the system" is typically underidentified.

In the discrimination literature, the solution has traditionally been to either assume correct beliefs (thus ascribing biased decisions to preferences) or model beliefs parametrically (or "statistically") and only ascribe the residual component of choices to preferences. Both approaches are controversial and few agreed-upon conclusions have been reached, despite the size of the literature on the topic (Guryan and Charles, 2013; Bertrand and Duflo, 2017; Neumark, 2018). Recent work has made clear more directly both that beliefs are often incorrect and non-Bayesian (see e.g. Lang and Lehmann, 2012a; Bohren et al., 2019; Bohren et al., 2025), and that allowing for differences across decision-makers in overall goals or knowledge further complicates inference (Canay et al., 2023).

We take a different approach. We focus on organizational contexts in which the overall goal relative to which agents' productivity (or, conversely, bias) must be measured is defined by law, and where their information set is observed. Such organizations are common in the public sector. The specific case we study is an anti-corruption agency's pursuit of indictments through whistleblower reports, some of which report on majority-group citizens (Peruvians of European descent) and others on minority-group citizens (Peruvians of indigenous descent). We present a theoretical framework and a corresponding experimental research design that allow identification not only of canonical taste-based and accurate statistical bias (Becker, 1971; Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977), but also incorrect beliefs-based bias and the researcher error terms that quantify the erroneous conclusions expected when correct beliefs are assumed. Our approach yields an overidentification test that additionally allows identification of incoherent agents—those whose actions cannot be rationalized by the econometrician. We show how to use a bias decomposition like ours that can be implemented at agent, team, or department-level to design policy that maximizes the organization's overall goal subject to the principal's bias rules and constraints, and how much realized bias against each group to expect.

Our approach has three core components: focusing on continuous rather than binary choices; directly measuring agents' beliefs about (also) observed downstream outcomes; and experimen-

tally varying the ethnicity agents see on some cases so as to "legally" measure the protected trait (here, "ethnicity itself"). Intuitively, our decomposition is the output of a form of sequential benchmarking (actions against beliefs, beliefs against reality) inspired by the way in which traditional marginal outcome tests benchmark actions against reality (see among many others Arnold et al., 2018; Arnold et al., 2022; Feigenberg and Miller, 2022).[1] Focusing on continuous actions—*priority scores* assigned to each case, analogously to *belief scores*—in effect makes every evaluated case marginal so that we avoid reliance on the IV assumptions that often complicate inference with marginal outcome tests. Equally important is that we identify not only taste-based bias but also incorrect beliefs. Assuming correct beliefs would notably flip the sign on estimated taste-based bias in the anti-corruption setting we study, and our framework shows that—and why—the same might well occur in many other settings too.

Quantifying ethnic bias in the Peruvian state's handling of whistleblower reports is an especially suitable application of the general framework and methodology we develop. First, corruption accusations are the central axis around which Peruvian politics, public trust, and policy debates tend to revolve. Second, the stakes are high. The grievances of Aymara, Quechua and other indigenous Peruvians dramatically burst into the open in the run-up to and aftermath of the 2021 presidential election.[2] Fifty to sixty percent of Peruvians consistently report believing that its indigenous minority—a quarter of the population—is discriminated against in day-to-day life (Defensoría del Pueblo, 2019). Third, "production" at the *Contraloría General de la República del Perú* (CGR) is organizationally similar to many state agencies across the world (see also Baron et al., 2024). Analysts evaluate whistleblower complaints in a two-step process: first, they decide whether a case should be *escalated*, that is, advanced internally for further investigation; second, the further and deeper downstream investigation of escalated cases determines if they ultimately result in an official *indictment*. A final feature of organizational settings such as the one we study that facilitate inference is that agents' information set is well-defined and observable: the case file (in our case whistleblower reports).

Consider a generic decision-maker who takes an action on the basis of their (potentially inaccurate) beliefs h. Belief-based bias—including accurate statistical discrimination as a special case—occurs when these beliefs depend directly on the subject's ethnicity (or, equivalently for this paper's purposes, race) r, and taste-based bias occurs when the *mapping* between beliefs and actions—which we denote ϕ_r—depends directly on race r. Our framework nests the case of a binary action, in which existing work has estimated taste-based bias through marginal outcome tests (MOT). Infra-marginal cases confound bias estimates so researchers typically study settings with quasi-random assignment of cases and variation in agents' leniency so that IV approaches pinpoint the marginal cases that identify bias if their assumptions hold (Chyn et al.,

[1] We compare our bias decomposition that uses observed beliefs to one that would have arisen if beliefs were erroneously assumed to be accurate for final outcomes. We show formally that the differences between these two are driven by how the extent of incorrect beliefs varies by the ethnicity of the case—itself a novel finding.

[2] Pedro Castillo, who is himself of indigenous origin, was seen to represent indigenous Peruvians and won the election. He was accused of corruption, impeached, and removed from office 16 months after taking office. Political instability has since increased.

2025; Goldsmith-Pinkham et al., 2026). With continuous decisions, every case is "marginal" for some level of the action, so infra-marginality bias is avoided quite simply by comparing cases that differ by ethnicity but share the same priority score. This greatly simplifies estimation.

The model also enables *testing* a relatively strong rationalizability restriction on decision-makers' choice rule that is needed to estimate bias using observational approaches: the so-called Extended Roy Model assumption (ERM) (Canay et al., 2023).[3] Our approach introduces and then imposes a relaxed version of ERM, which assumes only that there exist some characteristics of cases that are excluded from the function $\phi_r(\cdot)$, but still affect beliefs h. This notion of ERM is defined in the context of a general action space. We show that our weakened version of ERM yields a testable implication when beliefs h can be measured. We implement a statistical test for this implication of ERM, and find that the behavior of most—but not all—CGR analysts are compatible with the ERM restriction.

To measure priority and belief scores, 182 of CGR's 204 analysts participated in a triage experiment presenting them with real, concluded cases.[4] They were shown the same information set they would normally have access to—the case file—and asked to evaluate cases as in their day-to-day work. The only salient procedural differences were that agents were instructed not to keep their beliefs to themselves, nor to consolidate their preferred prioritization into a binary escalation recommendation. Instead they reported both—"how likely do you believe this case is to be indicted if escalated" and "how highly should this case be prioritized for escalation"—on a 0-100 scale. We vary the implied ethnicity of defendants in a randomly chosen subset of assignments as in audit studies (Bertrand and Mullainathan, 2004). This serves two purposes: first, experimental variation in race gives us a legally defined measure of the trait; and second, agents' response to such variation provides a measure of *overall* differential treatment of (and beliefs about) the two groups from which to start our decomposition (see also Bohren et al., 2024).

Changing the apparent race of a case will generally affect both h and ϕ. An additive decomposition requires a definition of each component of bias on a common scale. We achieve this by introducing a concept of "belief-equivalent bias". Belief equivalent bias captures the adjustment to the beliefs that a decision-maker has about a particular case from one ethnic group which would move the decision-maker to the same margin of action as a case from the opposite group that is otherwise observationally equivalent. Differences in belief-equivalent bias across races

[3]Canay et al. (2023) argue that (marginal) outcome tests of discrimination are prone to type-1 and type-2 errors because they implicitly assume homogeneous underlying risk distributions and agent decision thresholds, and that IV methods for identifying marginal cases are unlikely to overcome infra-marginality bias in more realistic scenarios. Instead they favor "fully" modeling agents' objective function and decision-making. Canay et al. (2023)'s approach indicates that (what they label an) Extended Roy Model (ERM) assumption is needed for outcome tests to identify (beyond-accurate-statistical) bias: that agents base choices on a single scalar latent index (expected gain or risk) and apply a (possibly group-specific) cutoff rule, so that selection is monotonic in that index and potential outcomes are determined by it. Hull (2021) argues that appropriate outcome tests are likely to more generally identify an observationally equivalent, as-if notion of (beyond-accurate-statistical) biased behavior, even if the econometrician cannot necessarily determine agents' true, underlying motivations.

[4]The experiment took place in July and August 2021. We ensured that no analysts saw cases from their own regional office in order to prevent familiarity with the actual details of the case and its outcome.

capture both taste-based and belief-based sources of bias, with the first component recovering the traditional estimand in the case of a binary action. In our audit study setting with a continuous action space, we show that belief-equivalent bias is identified from the experimental difference-in-means estimand after a particular transformation of the observed priority scores. An alternative transformation captures belief-based bias alone, and we can thus separate this bias from taste-based bias.

The sequential benchmarking intuition is simple. First, accurate statistical bias reflects how much higher (or lower) a particular case's true probability of downstream corruption indictment (its "risk") is if the accused is of indigenous rather than European ethnicity; we estimate this through a random forest prediction procedure with high goodness-of-fit. Second, belief-based bias reflects how much *more or less* than the true risk difference the agent's expected indictment probability responds to ethnicity switching. Finally, taste-based bias reflects how much *more or less* than the latter—the agent's true indictment beliefs—their preferred prioritization responds to ethnicity. We note that future research can extend our sequential benchmarking approach (even) further upstream or downstream in citizens' lifetime encounters with the state.[5]

In sum Peruvian anti-corruption agents appear remarkably biased against accused citizens of indigenous ethnicity. Our estimates suggest that analysts systematically overestimate the probability of indictment, but they do so more for indigenous defendants than they do for European defendents. Belief-based bias against indigenous cases is in fact of the opposite sign than the statistical discrimination that analysts would employ if they had accurate beliefs. Incorrectly assuming accurate beliefs attributes this differential to belief-based bias, leading to dramatically misleading estimates for both belief-based and taste-based bias. Sample-wide, we estimate taste-based bias against indigenous defendents that is equivalent to an increase of about two points in probability of "guilt", and total belief-based bias against Europeans. However, this belief-based bias is much smaller than it would be if analysts had accurate beliefs, which would lean strongly against European defendants.

In the final part of the paper we illustrate the direct mapping from the estimates coming out of our model to policy. With few exceptions (see especially Kline et al., 2024), the estimates from existing methods social scientists use to measure bias in the state are difficult to use in policy design. One reason is that they tend to conflate belief- and taste-based bias, which are widely

[5]Bias is generally speaking only defined relative to a downstream (or upstream) outcome. This can complicate interpretation of research findings when the benchmark outcome might itself plausibly be influenced by bias (such as when for example (potentially biased) police officers influence whether a defendant who was let out on bail by (potentially biased) judges and traditional marginal outcome tests are used characterize judge biases). It also, however, is an immutable and important feature of society that citizens "encounter" the state sequentially and that all such encounters involve human decisions and thus inherent scope for bias. Our model and estimates of (upstream) bias are well-defined and interpretable also when downstream outcomes are biased (as is also true in existing research making use of outcome tests). "Breaking up" the benchmark relative to which decisions are judged into (i) beliefs and (ii) actual downstream outcomes, however, arguably facilitates a cleaner interpretation if downstream outcomes are themselves biased. Our approach can easily be extended for example to additionally study (downstream) indictment bias, by benchmarking indictment decisions against beliefs about (further downstream) court outcomes and (such) beliefs against actual court outcomes. Correlation between upstream and downstream bias may represent a natural way to "micro-found" the study of systemic discrimination.

believed to require different policy responses. Another equally important reason is that existing methods typically need sample-wide variation to estimate biases with real-world data and thus cannot do so at agent, team, or department-level.[6] At the CGR, taste-based bias for example varies from 10 points against European defendants to nearly 30 points against indigenous defendants. We demonstrate the implications for a state agency with different political or ethical considerations constraining the organization's mission by focusing on a particular policy lever: assignment of agents to cases. A *Contraloría* which maximizes indictments subject to minimal, realized taste-based bias would for example assign agents to cases very differently from one which does so subject to minimal taste-based *and* belief-based bias, and we predict markedly different realized bias against indigenous and European-origin Peruvians in the two cases.

This paper is a forward-looking, policy-oriented attempt to bridge and aggregate what economists studying biased decision-making have learned the last few decades and take seriously both that beliefs are often incorrect and themselves biased and that some behaviors aren't (yet, at least) rationalizable. Our model and corresponding experimental methodology is to our knowledge the first attempt at decomposing overall differences in how (state) agents treat majority and minority citizens into the different forms of biases documented in the literature in a way that can be used in policy design. Our analysis thus relates to work on anti-discrimination policy design; on methods for measuring bias; and on concrete ways to increase state effectiveness. [FULL LITERATURE REVIEW AND CITATIONS TBA!]

[6]Some existing work making use of outcome tests to estimate bias consider more than one group of agents or citizens (see e.g. Mueller-Smith, 2015). While useful for reducing the likelihood that non-monotonic decision-making renders IV estimates invalid, this approach is inherently limited for policy design. We follow Kline et al. (2022) and Kline et al. (2024)'s estimates of different U.S. firms' discrimination propensities in attempting to directly inform policy by characterizing the behavior of individual "units".

2 Context, Data and Experimental Design

2.1 Context: Whistleblower Complaints of Corruption in Peru

This study is set in Peru, where corruption is a serious problem: it ranks as the 52nd most corrupt country globally and 51.6% of firms identify corruption as a major or severe constraint (Transparency International, 2026; Bank, 2023). Yet civil society is remarkably active in denouncing acts of corruption. A key mechanism through which citizens can expose corruption is through making whistleblower complaints to the comptroller general's office, *Contraloría General de la República* (hereafter CGR).

Any citizen can make a complaint to CGR through a variety of channels (online, by phone, in person at a CGR office, or at the regular town-hall meetings CGR holds throughout the country). Once a complaint is made, it enters CGR's system, and a team of approximately 200 civil servants processes it in three steps.

First, the whistleblower complaint goes through a preliminary *screening* step to ensure that basic criteria are met. Specifically, the case must include a detailed description of the irregular activity, the government entity in which it occurred, the relevant date or time-frame, and the identity of any public official implicated by the whistleblower. It must describe alleged acts or omissions by public officials that are contrary to applicable regulations or provisions (such as legal rules, internal policies, contractual stipulations, among others).

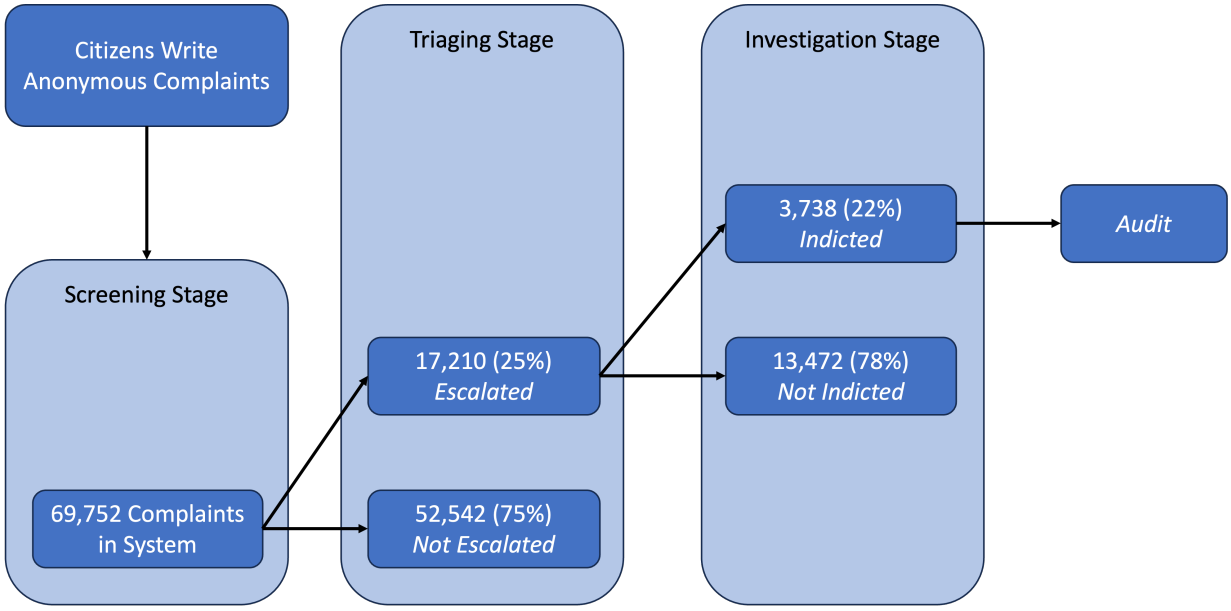
If the case passes this initial screen, it proceeds to the second, *triaging* stage. An analyst reads the case alongside any accompanying evidence. The analyst may determine that CGR lacks jurisdiction.[7] If the analyst determines that the complaint merits an in-depth investigation, they recommend that it be *escalated*. Otherwise, it is dismissed.

Finally, escalated cases advance to the *investigation* stage, where analysts examine the evidence provided, supplemented by any additional evidence they may request from the accused public agency. Based on this in-depth investigation, the analyst then either rejects the case (for example, if the evidence is not specific or strong enough) or issues an *indictment* recommending a full audit of the public entity. In the event of an indictment, the case is handed over to CGR's auditors who take the case forward. After that, CGR issues recommendations to the public entity for improvements in its procedures as well as potentially issuing administrative sanctions on the officials involved.

Between 2014 and 2021, CGR received an average of 8,720 complaints per year. On average, 458 cases were cleared per year, leading to an accumulating backlog. Analysts may face pressure to make quick screening and triage decisions, potentially making them vulnerable to implicit biases (Dror, 2020; Kassam, Kosseff, & Green, 2009). Figure **??** shows the lifecycle of the 69,752 complaints received between 2014 and 2021. It shows that 17,210 were admitted for further evaluation in the triaging stage, while 52,542 were rejected, and those that were addressed

[7] Based on Law No. 27785, CGR is not responsible for handling complaints that do not involve public resources. For instance, common crimes, labor disputes, or issues with private entities that do not concern public funds are referred to the appropriate regulatory authority.

FIGURE 1: WHISTLEBLOWER COMPLAINT LIFECYCLE



were handled under resource constraints. This represents 25% of the complaints submitted by citizens. Subsequently, of the 17,210 escalated complaints, 3,738 (22%) led to an indictment and a recommendation to audit the implicated public entity.

2.2 Administrative Data

Through our partnership with CGR, we are able to work with administrative data on the full universe of whistleblower complaints (as summarized in figure 1). The data contains rich records on the details of each case and the decisions made by analysts and their supervisors at each stage of the complaint's lifecycle. In their initial submissions, citizens very often mention the name of the public official they are accusing of corruption. Based on these names, we assign each named defendant an ethnicity inferred by the ethnic origin of their surname. We rely on ethnographic work by Artiles (2024) to trace the linguistic roots of surnames in Peru.[8]

The richness of the administrative data allows us to form remarkably accurate predictions of the decisions that will be taken at each stage, as detailed in appendix D. Moreover, by training our classification algorithms using data on the ethnicity of the defendant, we are able to estimate the indictment probability of each case both with the ethnicity that the defendant actually has and also assuming that the defendant had the opposite ethnicity. Comparing the difference in these predictions to the differences in beliefs we elicit through the experiment allows us to estimate the extent of inaccurate belief bias (as argued in section 3.3).

[8]We are immensely grateful to Miriam Artiles for sharing this dictionary with us.

2.3 Experimental Design

Our experiment was designed to emulate the escalation decisions that analysts make in the Triaging stage of processing a complaint. In July and August 2021 we held 14 sessions with groups of between 10 and 18 CGR analysts in which a total of 182 analysts (approximately 90% of the nationwide triaging staff) reviewed real cases. For each case, the analysts were asked to carefully read the complaint and then provide their assessment of the case in two ways. First, analysts assigned the case a *priority* score from 0–100 indicating the priority they would give the case for escalation. Second, analysts assigned each case an *indictment probability* score from 0–100 indicating their belief that the case would eventually lead to an indictment. Appendix figure C.1 shows the screens analysts saw displaying the text of the complaint and eliciting their two scores. Participants were given 10 trial rounds to practice using the 0–100 slider. They were also allowed to go back to previous complaints and update their scores at any point during the experiment, although in practice less than 2% of scores were changed.

For the experiment, we created a library of 196 real cases. The cases were selected to satisfy five eligibility requirements: (i) text length between 50 and 350 words; (ii) issued between 2018 and 2020; (iii) allegations involving illicit enrichment, favoritism, or deficient public works execution; (iv) defendants identified as municipal mayors; and (v) defendant's full name explicitly mentioned. The final sample aimed to reflect Peru's ethnic composition.[9] On average, each analyst reviewed 105 unique complaints.

Randomization Procedure. To experimentally manipulate the apparent ethnicity of a case, we randomly switched defendant surnames between Indigenous (Quechua and Aymara) and European (Spanish) origins using regional surname dictionaries derived from historical CGR data and Artiles (2024).[10] The randomization was stratified by defendant ethnicity (Indigenous or European) and case outcome (dismissed at escalation, escalated but dismissed at indictment, or indicted). About a quarter of cases had their surnames switched. Analysts were then assigned a random sequence of cases with the same stratification and ensuring that no analysts saw cases from their own regional office to prevent familiarity with the cases.[11]

3 Theory: bias in single-phase decisions

3.1 Basic model

We model bias in a setting in which a decision-maker j faces a a single decision in which they choose an action $a \in \mathcal{A}$ after observing r and x. Here r denotes an identity characteristic (e.g.

[9] In the 2017 Peruvian Census, around 25% of the population identified as either Quechua or Aymara.

[10] Complaints in the experimental sample came from every regional office in Peru. When determining the dictionary of surnames to use for randomly switching we took into considerations the most common last names in the region that the complaint came from.

[11] Analysts only have jurisdiction over complaints coming from their respective regional office. The detailed randomization algorithm is presented in appendix G.

"race") of the person that the complaint is about, and x all other variables regarding the complaint. We refer to the decision-maker as an *analyst* and the person that the complaint is about as the *accused*. We refer to the complaint as a "case".

We assume that decisions a for a given analyst are driven by a scalar index variable $h_j(r, x)$ that can depend generally on case observables (r, x), as well as individual aspects of the analyst j (e.g. individual knowledge, cognitive biases, etc.). We refer to $h_j(r, x)$ as analyst j's *belief* about case (r, x). Decisions can also depend arbitrarily on a strict subset of the observables x that we denote as v. Components of x that are not included in v only affect actions through $h_j(r, x)$. Following Canay et al. (2023) we refer to this restriction as an "Extended Roy Model" or (ERM).

Our notion of an ERM weakens the restriction considered by Canay et al. (2023) by allowing for the function ϕ to depend on a subset of the x through v. Formally, we suppose that

$$a = \phi_{jrv}(h_j(r, x)) \tag{1}$$

for some increasing and right-continuous function ϕ_{jrv}.[12]

Example 1: binary actions. Most of the existing literature on the econometric evaluation of discrimination considers binary actions $\mathcal{A} = \{0, 1\}$. For binary actions, ϕ_{jrv} cannot be a *strictly* increasing function (provided that $h_j(r, x)$ takes on more than two values). In our analysis of the historical escalation decision, we accordingly assume ϕ_{jrv} only to be *weakly* increasing. In the binary case, we can without loss of generality write $\phi_{jrv}(h) = \mathbb{1}(h \geq t_{jrv})$, where t_{jrv} represents a threshold level of belief at which j takes the action labeled 1, for a case with observables (r, v).[13]

Example 2: continuous actions. When $\mathcal{A}$ is an ordered continuum, such as in our lab experiment for the priority score, assume instead that ϕ_{jrv} is *strictly* increasing. The intuitive meaning is that if j's belief about case increases, their action must increase as well.

Remark: Eq. (1) can be derived from a utility-maximization model in which $h_j(r, x)$ represents analyst j's subjective evaluation that a given decision will result in an indictment, or alternatively that the defendant is "guilty", indicated by a binary random variable Y^*. That is, $h_j(r, x) = \mathbb{E}_j[Y^*|R = r, X = x]$ where $\mathbb{E}_j[\cdot|\mathcal{I}]$ indicates analyst-j's (potentially inaccurate) subjective beliefs given information $\mathcal{I}$. In this case, suppose that j choses a to maximize

$$\mathbb{E}_j[u_j(Y^*, a, r, v)|R = r, X = x] = u_j(0, a, r, v) + h_j(r, x) \cdot \{u_j(1, a, r, v) - u_j(0, a, r, v)\}$$

[12] Right-continuity of ϕ_{jrv} is essentially a normalization: by Lebesque's theorem for monotone functions, ϕ_{jrv} can only have isolated points at which it is discontinous. At the countable set of values h that correspond to these discontinuities, we assume $\phi_{jrv}(h)$ to be equal to its left limit at h.

[13] Existing studies consider a binary action, e.g. pre-trial release, that is taken when beliefs (about guilt or misconduct potential) are *low*. For example , the models of Arnold et al. (2018) and Canay et al. (2023) can be written as $a = \mathbb{1}(h_j(r, x) \leq t_{jrv})$, where $a = 1$ indicates release. This can be adapted to the form of Eq. (1) by redefining $a = 1$ to be non-release (while changing the weak inequality into a strict inequality to yield right-continuity rather than left-continuity of the resulting increasing function).

over $a \in \mathcal{A}$. The maximizing a of expected utility, assuming it is unique, can from the above be written without loss of generality as $\phi_{jrv}(h_j(r,x))$. For $\phi_{jrv}(\cdot)$ to be weakly-increasing, it is sufficient that $\{u_j(1,a,r,v) - u_j(0,a,r,v)\}$ be weakly increasing in a:[14] for example if the analyst suffers a utility loss $\{u_j(1,a,r,v) - u_j(0,a,r,v)\} < 0$ from the recipient of the decision being guilty rather than innocent, the magnitude of this loss is reduced if the analyst chooses a higher action a. This is an intuitive property if a is a punishment or mitigating action such as conviction, investigation, or incarceration.

3.2 Defining belief-based and taste-based bias

Consider two values of r, which we denote as e and i for European and Indigenous, respectively.

We say that *belief-based bias* (BBB) holds for the case-analyst pair (j,r,x) when $h_j(i,x) \neq h_j(e,x)$: that is beliefs would differ between two cases with different races but otherwise identical on observables x. We say that *taste-based-bias* (TBB) holds when the mapping between these beliefs and actions is different: $\phi_{jix} \neq \phi_{jex}$, i.e. the analyst would treat Indigenous and European cases differently even if they shared the same belief about them.

To measure the magnitudes of TBB and BBB both on a common scale, we introduce the concept of "belief-equivalent" total bias. Belief-equivalent bias considers the following thought experiment: what adjustment to the analyst's belief about a European such that they would choose the same action as they would for an Indigenous case that otherwise has the same observables? This affords a measure of bias in units of beliefs, but incorporating bias that comes both from beliefs as well as from preferences. Contemplating the "smallest" such adjustment, we define:

$$\delta_{jex} := \sup\{\delta : \phi_{jev}(h_j(e,x) + \delta) \leq \phi_{jiv}(h_j(i,x))\} \tag{2}$$

δ_{jex} is the adjustment to j's belief that would just push them to give a European case with non-race observables x at least as a high an action than an Indigenous case with these same observables. δ_{jex} is strictly positive when the effect of changing r from e to i on the action taken by j given x: $\phi_{jiv}(h_j(i,x)) - \phi_{jev}(h_j(e,x))$, is strictly positive, allowing a strictly positive shift to beliefs δ for a European without leading to a larger action than the one taken against an observationally-equivalent Indigenous case. The condition that $\phi_{jiv}(h_j(i,x)) > \phi_{jev}(h_j(e,x))$ can occur when there is TBB against Indigenous cases: $\phi_{jiv}(h) > \phi_{jev}(h)$ for a given h, when there is BBB against Indigenous cases: $h_j(i,x) > h_j(e,x)$, or a combination of both.

We can similarly define an alternative measure of belief-equivalent bias that puts δ within the ϕ_{jiv} function rather than the ϕ_{jev} function:

$$\delta_{jix} := \sup\{\delta : \phi_{jev}(h_j(e,x)) \geq \phi_{jiv}(h_j(i,x) + \delta)\} \tag{3}$$

[14] In this case the function $f_{jrv}(a,h) := u_j(0,a,r,v) + h \cdot \{u_j(1,a,r,v) - u_j(0,a,r,v)\}$ satisfies the property of *increasing differences*, i.e. that for $a' \geq a$, $f_{jrv}(a',h) - f_{jrv}(a,h)$ is monotone decreasing in h (Milgrom and Shannon, 1994).

The quantity δ_{jix} is the adjustment to j s beliefs about an Indigenous case with non-race observables x that would just push it over the edge of delivering as high an action than for an observationally-equivalent European case. We have $\delta_{jix} > 0$ when $\phi_{jiv}(h_j(i,x)) < \phi_{jev}(h_j(e,x))$, in exactly the opposite situation as when δ_{jex} is positive. We can thus define a symmetric notion of the total belief-equivalent bias against Indigenous cases from the difference:

$$S\text{-}TOT_{jx} := \frac{1}{2}(\delta_{jex} - \delta_{jrx})$$

where the prefix S- emphasizes the combination of δ_{jex} and δ_{jrx} to obtain a symmetric measure of total discrimination. In Section **??**, we focus on the identification of δ_{jex} across collections of analysts J and cases X—rather than the symmetric aggregate $S\text{-}TOT$—which substantially simplifies the exposition. However, below we show that $S\text{-}TOT$ more elegantly unifies our definitions with existing work in the case of a binary action space.[15]

We can obtain an explicit expression for δ_{jix}, δ_{jix} (and hence $S\text{-}TOT_{jx}$) by defining $\phi^{-1}_{jrv}(a) := \inf\{h \in \mathbb{R} : \phi_{jrv}(h) > a\}$. In the case of continuous actions, ϕ^{-1}_{jrv} is simply the inverse of the strictly increasing function ϕ_{jrv}. If ϕ_{jrv} is only weakly increasing, then ϕ^{-1}_{jrv} is a generalized inverse function of ϕ_{jrv} that maintains many of the same properties.

Proposition 1. *$\delta_{jex} = \phi^{-1}_{jev}(\phi_{jiv}(h_j(i,x))) - h_j(e,x)$ and $\delta_{jix} = \phi^{-1}_{jiv}(\phi_{jev}(h_j(e,x))) - h_j(i,x)$.*

Proposition 1 holds regardless of whether $\mathcal{A}$ is binary, discrete, or continuous. As a consequence of Proposition 1, we have that:

$$S\text{-}TOT_{jx} = \frac{1}{2}\{\phi^{-1}_{jev}(\phi_{jiv}(h_j(i,x))) - \phi^{-1}_{jiv}(\phi_{jev}(h_j(e,x))) + h_j(i,x) - h_j(e,x)\} \tag{4}$$

Expression (4) is useful in showing how $S\text{-}TOT_{jx}$ combines TBB and BBB. We see this by considering the binary and continuous cases each in turn.

Example 1: binary actions. Suppose that $\phi_{jrv}(h) = \mathbb{1}(h \geq t_{jrv})$. Then $\phi^{-1}_{jrv}(0) = t_{jrv}$ and $\phi^{-1}_{jrv}(1) = \infty$. Suppose that $\phi_{jiv}(h_j(i,x)) = \phi_{jev}(h_j(e,x)) = 0$ so that δ_{jex} and δ_{jix} are both well-defined. Then $S\text{-}TOT_{jx}$ provides an average of bias in tastes and bias in beliefs:

$$S\text{-}TOT_{jx} = \frac{1}{2}\underbrace{\{t_{jev} - t_{jiv}\}}_{\text{taste-based bias}} + \frac{1}{2}\underbrace{\{h_j(i,x) - h_j(e,x)\}}_{\text{belief-based bias}},$$

Notice that each term is measured on a comparable scale of beliefs.[16] A describes how as-good-as-random assignment to analysts can be used to isolate an average of $t_{jev} - t_{jiv}$ across judges.

[15] When the functions ϕ are approximately linear and with a similar slope for Indigenous and European cases, i.e. $\phi_{jrv}(h) \approx \alpha_{jrv} + \gamma_{jv} \cdot h$ for some α and γ, then one can show that $\delta_{jex} \approx -\delta_{jix}$. This approximation appears to hold in our setting, making δ_{jex} and $S\text{-}TOT_{jx}$ very similar quantitatively.

[16] If instead $\phi_{jiv}(h_j(i,x)) = \phi_{jev}(h_j(e,x)) = 1$ and we replace ∞ with a large fixed number $\bar{H}$ rather infinity in the definition of ϕ^{-1}_{jrv}, then the $\bar{H}$ cancel out and $TOT_{jx} = 1/2 \cdot (h_j(i,x) - h_j(e,x))$ captures belief-based bias alone.

Example 2: continuous actions. Consider the expression for δ_{jex} given in Proposition 1. By adding and subtracting $h_j(i,x)$, we can rewrite it as:

$$\begin{aligned}\delta_{jex} &= \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x) \\ &= \underbrace{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x)}_{\text{taste-based bias}} + \underbrace{h_j(i,x) - h_j(e,x)}_{\text{belief-based bias}} \end{aligned} \tag{5}$$

When ϕ_{jev} is invertible, the first term can only differ from zero if $\phi_{jev} \neq \phi_{jiv}$ as a function. Meanwhile $h_j(i,x) - h_j(e,x)$ captures BBB alone. Thus δ_{jex} captures a sum of TBB and BBB. By similar steps one sees that $\delta_{jix} = \underbrace{\phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x))) - h_j(e,x)}_{\text{taste-based bias}} - \underbrace{\{h_j(i,x) - h_j(e,x)\}}_{\text{belief-based bias}}$ and thus altogether:

$$S\text{-}TOT_{jx} = \underbrace{\frac{1}{2}\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x)\} + \frac{1}{2}\{h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\}}_{\text{taste-based bias}} + \underbrace{h_j(i,x) - h_j(e,x)}_{\text{belief-based bias}}$$

Thus we have seen both in the binary and in the continuous action settings $S\text{-}TOT_{jx}$ captures a linear combination of TBB and BBB: in the continuous action case its two components δ_{jrx} also have this interpretation for each $j \in \{i,e\}$.[17]

3.3 Separating out the role of incorrect beliefs

While the difference $BBB_{jx} := h_j(i,x) - h_j(e,x)$ captures belief-based bias, let $ASD_{jx} := h(i,x) - h(e,x)$ represent the difference in *true* indictment probabilities between Indigenous and European cases that share the same value x of non-race observables. This difference is commonly referred to as *accurate statistical discrimination*. ASD differs from belief-based bias when analysts have incorrect beliefs about the probability that a given case will result in indictment, i.e. when $h_j(r,x) \neq h(r,x)$. To this end, we can write write

$$BBB_{jx} = ASD_{jx} + IBB_{jx}$$

where $IBB_{jx} = \{h_J(i,x) - h(i,x)\} - \{h_J(e,x) - h(e,x)\}$. Belief-based bias will differ from accurate statistical discrimination when analysts are "more wrong" about cases of one race than the other, and thus $IBB \neq 0$.

Thus, we can decompose total belief-equivalent bias into three terms, one arising from preferences, one arising from accurate statistical discrimination, and one from incorrect beliefs:

[17] By contrast, in the binary action case note that for example when $\phi_{jiv}(h_j(i,x)) = 0$ we have that $\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x) = t_{jev} - h_j(e,x)$ which is not interpretable as a measure of TBB: it can differ from zero even if $t_{jev} = t_{jiv}$. It is for this reason that we must appeal to the symmetric measure $S\text{-}TOT_{jx}$ of total belief based bias directly in the binary action case in order to see a decomposition into TBB and BBB, while we can decompose each term δ_{jrx} separately in the continuous case.

$$S\text{-}TOT_{jx} = TBB_{jx} + ASD_{jx} + IBB_{jx}$$

4 Identification and estimation

In this section we provide a set of assumptions under which we are able to identify each of the terms in the decomposition in section 3.3 using our triage experiment and administrative data to implement the decomposition. In section 4.1 we develop notation to link the theory in section 3 to the data and define the empirical analogs of the terms in the decomposition in 3.3 that we will seek to estimate. In section 4.2 we provide the assumptions under which we can identify each of the terms in our decomposition and state our main identification result.

The key substantive assumption about analysts' decisions that permits the decomposition is our Extended Roy Model (ERM) assumption. In section 4.3 we show that this assumption has testable implications that we test in our data. Implementing these tests, we find that between 10 and 20% of analysts fail to satisfy the ERM restriction. Finally in section 4.5 we highlight the value of direct data on beliefs in permitting our decomposition and we quantify the bias that would arise from attempting a decomposition using only data on analyst choices and final outcomes, as previous work has tended to do (CITES).

4.1 Notation and estimands

The unit of analysis is a pair (c, j), where c denotes a real-world case and j an analyst in the triage experiment. We refer to such a pair as a "case-pair", and analysts in the triage experiment as experiment-analysts. We use capital letters without indices to denote random variables defined over a population of *case-pairs* in the triage experiment. We have eight primary variables of interest are (R, X, Y, S, H, J, F, B).

Each case c has observables (R_c, X_c, Y_c), where (R_c, X_c) is the information available to the lab-analyst and Y_c is an indicator for whether the case actually resulted in an indictment (unknown to the lab analyst). During the triage experiment, each case-pair is randomly assigned an experiment-analyst J who reads the case and assigns it a priority score $S_{c,J}$, and subjective likelihood of indictment $H_{c,J}$.

As discussed in section EXPERIMENT, within blocks of cases denoted by B, we randomly select case-pairs to be shown to the experiment-analyst "flipped". F indicates whether the ethnicity of the case was flipped before showing it to the lab-analyst J in that case-pair, and B are blocks of case-pairs within which flips were randomly assigned.

We let $\tilde{R}$ denote the "apparent" race of a case by the lab-analyst, where

$$\tilde{R} = \begin{cases} e & \text{if } R = e \text{ and } F = 0 \\ i & \text{if } R = e \text{ and } F = 1 \\ i & \text{if } R = i \text{ and } F = 0 \\ e & \text{if } R = i \text{ and } F = 1 \end{cases}$$

where $R \in \{i, e\}$ are shorthand for Indigenous and European, respectively.

As apparent race $\tilde{R}$ is what is manipulated experimentally, we introduce potential outcomes for our key outcome variables in the triage experiment. Let $s_j(r, x)$ denote the priority score that analyst j would assign to a case with observables (r, x), and thus $S = s_J(\tilde{R}, X)$. Similarly, let $h_j(r, x)$ be potential outcomes for an analysts' subjective likelihood of indictment.

We do not assume that realized subjective likelihoods $h_J(\tilde{R}, X)$ are directly observable. Rather, we allow each case-pair to have a realization U of measurement error in the observed subjective likelihood scores o that $H = H^* + U$ where $H^* = h_J(\tilde{R}, X)$. We do not allow for such measurement error in scores S, as S represents the action "actually" taken by the lab-analyst in the lab experiment. The existence of U allows for H to have a non-degenerate distribution conditional on S (and $\tilde{R}, J, V$), and vice-versa. Such non-degeneracy is observed in our data.

Parameters of interest for estimation. In section 3 we defined the decomposition of total bias into its three constituent terms conceptually. We now convert these into empirical averages that we will estimate from our data. In Section 3.2, we introduced the total belief-equivalent bias TOT_{jx} for analyst j and a case with observables x. We defined TOT_{jx} in a manner that was symmetric between the European and Indigenous categories, as an average of δ_{JeX} and δ_{JiX}. This allowed for TOT_{jx} to yield a particularly parsimonious expression in the case of binary actions, while remaining easily interpretable in the case of continuous actions as well. For our econometric analysis, we define our total bias parameter TOT as an average of the belief equivalent bias δ_{JeX}, rather than as an average of TOT_{JX}. This yields substantially shorter expressions and improves exposition.

We define average accurate statistical discrimination against Indigenous cases as

$$ASD := \mathbb{E}[h(i, X) - h(e, X)]$$

where $h(r, x) = \mathbb{E}[Y|R = r, X = x]$. ASD represents the average difference in true indictment probabilities between observationally equivalent Indigenous and European cases. Since the cases that appear in the lab experiment are drawn randomly from the population of historical cases (and each shown is to an approximately equal number of analysts), the expectation over case-pairs in the definition of h recovers the historical distribution over cases. This is important for the interpretation of h as an *accurate* measure of indictment probability.

Next, we define average belief based bias against Indigenous cases as

$$BBB := \mathbb{E}[h_J(i, X) - h_J(e, X)]$$

which averages over the distribution of analysts J and of case observables X in the triage experiment. BBB quantifies how much higher analysts' subjective beliefs about indictment probabilities are, on average, for an Indigenous case as compared with an otherwise observationally-equivalent European case.

BBB differs from ASD when analysts have incorrect beliefs about the probability that a given case will result in indictment, i.e. $h_j(r, x) \neq h(r, x)$. To this end write

$$BBB = ASD + IBB$$

where $IBB := \mathbb{E}[h_J(i, X) - h(i, X)] - \mathbb{E}[h_J(e, X) - h(e, X)]$. Belief-based bias will differ from accurate statistical discrimination whenever analysts hold beliefs that are further from accurate beliefs for cases of one race than the other, resulting in $IBB \neq 0$.[18] We refer to IBB as "incorrect belief bias" or "inaccurate statistical bias".

Taste based bias is defined in section 3.2 as occurring whenever the mapping between beliefs and actions is different for indigenous and european cases. Whenever taste based bias is present, two cases for which analysts hold the same beliefs lead to different actions. And conversely, two cases which lead to the same action must be driven by different beliefs. As discussed in section 3.2, we can measure the extent of taste-based bias by quantifying the difference in beliefs that would be required to generate the same actions for cases with different ethnicities.

We define average taste based bias against indigenous cases as

$$TBB := \mathbb{E}_S\big[\mathbb{E}[h_J(e, X)|s_J(e, X) = S] - \mathbb{E}[h_J(i, X)|s_J(i, X) = S]\big] \tag{6}$$

This taste based bias measures the difference in indictment probability beliefs between indigenous and European cases that are given the same priority score.

Recall from Section 3.2 that δ_{JeX} is the adjustment to J's belief about a European case with observables X that would cause them to take the same action as they would for an Indigenous case with observables X. Letting $TOT := \mathbb{E}[\delta_{JeX}]$ we have by equation (5) that:

$$TOT = TBB + ASD + IBB \tag{7}$$

which defines the estimands we will seek to estimate using our triage experiment and administrative data. Next, we provide assumptions under which these terms are all identified and discuss how we can estimate them.

[18] Note that if the set of cases in the lab experiment were not drawn at random from the population we could also have that $h_j(r, x) \neq h(r, x)$ even if analysts are "correct" in their beliefs about indictment probabilities in the broader population. In general, the quantity $\tilde{h}_j(r, x) := h_j(r, x) - h(r, x)$ will combine both of these effects.

4.2 Assumptions and identification

A decomposition of total bias into taste-based and belief-based components is feasible under weak assumptions given i) the continuous action space for the priority score; ii) the elicitation of beliefs about indictment probability H, and iii) the experimental flipping of apparent race using the audit study. We now formalize the required assumptions.

Our first and most important assumption for identification is that an analyst's subjective probability that the case will result in an indictment represents an index that drives their priority score decisions in an invertible manner.

Assumption: ERM. (Extended Roy Model with Continuous Actions). $s_j(r,x) = \phi_{jrv}(h_j(r,x))$ *for some strictly increasing function* ϕ_{jrv}*, where* v *is a strict subset of the observable components of* x*.*

Assumption ERM embeds a few key restrictions. Most substantively, the function ϕ_{jrv} does not depend on the full vector x. If v were instead allowed to be equal to the full set of observables x, then assumption ERM would hold trivially.[19] In the other extreme, if no variables are contained in v then Assumption ERM extends the notion from Canay et al. (2023) of an "extended Roy model" (ERM) to our setting in which the action space (priority scores) is continuous rather than binary. For intermediate v (embeding some but not all of the observables X), Assumption ERM allows the preferences of lab-analysts over values of the priority score to assign to depend on certain observables of the case-pair beyond race: for example the type of violation that the defendant is accused of. This is important if experiment-analysts see indictments for e.g. illicit enrichment as more important than e.g. favoritism.

Assumption ERM does not take a strong stand on what analysts are targeting: it is, for example, compatible with analysts attempting to maximize indictments (regardless of guilt), and is also compatible with analysts attempting to maximize indictments among cases in which the analyst feels that the accused is actually guilty, or maximizing indictments in which the accusation is "severe", etc. Whatever the analysts' objective function, ERM requires priority scores to be strictly increasing in the analyst's own perception of the likelihood of indictment, between cases in which case properties v are held fixed. Note that ERM allows for analyst-specific heterogeneity in tastes for priority score, though the set v is taken to be common over all analysts (intuitively, one can take v to be the union of the variables that any analysts care about).

In the triage experiment, we define randomization strata defined from case observables R and Y. Denote these by the random variable B, where $B = s(R,Y)$.[20] Let L be the regional office (or "location" of a given case). Our second assumption is that randomization was implemented successfully:

[19] In the case of $v = x$ ERM would amount to assuming that $s_j(r,x) = \phi_{jrx}(h_j(r,x))$ for some strictly increasing function ϕ that can depend directly on x. Then ERM would become trivial because whatever the functions h_j and s_j are, we could simply define ϕ_{jrx} as $\phi_{jrx}(h) = s_j(r,x) - h_j(r,x) + h$.

[20] In a slight abuse of notation, we here let Y in the funciton b denote which of *three* outcomes occured for a case: not escalated, escalated but not-indicted, or indicted. We otherwise let Y be a binary variable indicating whether or not the case was indicted. Further details on the construction of the strata can be found in Appendix G.

Assumption: RAND. (Randomization of Experiment-Analysts and Ethnicity Flips) *The following hold:*

1. $\{J \perp\!\!\!\perp (R, X, Y)\}|L$
2. $\{F \perp\!\!\!\perp (R, X, Y)\}|(B, J)$

where the first part of the assumption conditions on location L since we avoided showing any analyst cases from their own region. Our third assumption is that measurement error in subjective indictment probabilities is classical in the following sense:

Assumption: CLASSICAL. (Classical Measurement Error) $\mathbb{E}[U|J, B, X, \tilde{R}, R] = 0$

Assumption CLASSICAL permits an idiosyncratic, mean-zero gap between analysts' self-reported subjective indictment probability and the one that they actually use in making their decision of S. It holds provided that $\{U \perp\!\!\!\perp X\}|J, B, \tilde{R}, R$ with $\mathbb{E}[U|J, B, \tilde{R}, R] = 0$.

Identification With these assumptions we can show how we are able to identify each of the components of the decomposition in equation (7). We will assume that the researcher has identified a candidate set of observables v under which the researcher believes Assumption ERM to hold (and in section 4.3 we return to the choice of v through a testable implication of Assumption ERM). Define the conditional expectation

$$g_{jv}(r, s) := \mathbb{E}[H|R = r, S = s, J = j, V = v, F = 0] \tag{8}$$

The quantity $g_{jv}(r, s)$ denotes the expected subjective likelihood of indictment among (unflipped) cases of ethnicity r with $V = v$, for which analyst j assigns a priority score s.

The following Lemma proves useful in detecting taste-based bias:

Lemma 1. *Under ERM and CLASSICAL, the function $\phi_{jrv}(h)$ is identified from the data as $\phi^{-1}_{jrv}(s) = g_{jv}(r, s)$*

Remark: Lemma 1 extends the classic MOT from binary-action settings, in which the analog of g_{jv} is the probability of some downstream measure of guilt or misconduct (e.g. pre-trial misconduct in settings studying judges' bail decisions as in Arnold et al. (2018) and Arnold et al. (2022)). By Lemma 1, we see that an analog of the MOT in our continuous-action setting is simply to look at

$$g_{jv}(e, s) - g_{jv}(i, s) = \phi^{-1}_{jev}(s) - \phi^{-1}_{jiv}(s) \tag{9}$$

for any fixed value of s. Given that all cases with $S = s$ are *marginal* for having an action that is less than or equal to s, the LHS of (9) quantifies TBB against Indigenous defendents among cases at that margin s. For example, a positive value indicates that to give a priority score of s, agent j would have to have a higher belief about a European case's probability of indictment than they would for an Indigenous case.

This also clarifies that we can rewrite our estimand of taste-based bias (6) in terms of ϕ_{JrV} as

$$\begin{aligned} TBB :=& \mathbb{E}_S\left[\mathbb{E}[h_J(e,X)|s_J(e,X)=S] - \mathbb{E}[h_J(i,X)|s_J(i,X)=S]\right] \\ =& \mathbb{E}\left[\phi_{JeV}^{-1}\left(\phi_{JiV}\left(h_J(i,X)\right)\right) - h_J(i,X)\right] \end{aligned}$$

In using Lemma 1 to establish identification, it is useful to write our bias parameters in terms of disaggregated versions of that parameter which condition on cases with a given race r, e.g. $ASD_r := \mathbb{E}[h(i,X) - h(e,X)|R=r]$, $IBB_r := \mathbb{E}[\tilde{h}(i,X) - \tilde{h}(e,X)|R=r]$, $TBB_r := \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h_J(i,X))) - h_J(i,X)|R=r]$ and $TOT_r := \mathbb{E}[\delta_{JeX}|R=r]$. By the law of iterated expectations, the unconditional analogs can then be computed, e.g. $ASD = pASD_e + (1-p)ASD_i$ where $p = P(R=e)$ is observable.

Theorem 1. *Under ERM, RAND and CLASSICAL, we can identify the race-conditioned version of each component of Eq. (7) as follows:*

- $ASD_r = \mathbb{E}\left\{\mathbb{E}[Y|R=i,X] - \mathbb{E}[Y|R=e,X]|R=r\right\}$
- $IBB_r = BBB_r - ASD_r$
- $TBB_r = TOT_r - BBB_r$

where BBB_r and TOT_r are in turn identified as:

1. $BBB_e = \mathbb{E}\left\{\mathbb{E}[g_{JV}(i,S)|R=e,J,B,F=1] - \mathbb{E}[g_{JV}(e,S)|R=e,J,B,F=0] \mid R=e\right\}$
2. $TOT_e = \mathbb{E}\left\{\mathbb{E}[g_{JV}(e,S)|R=e,J,B,F=1] - \mathbb{E}[g_{JV}(e,S)|R=e,J,B,F=0] \mid R=e\right\}$
3. $BBB_i = \mathbb{E}\left\{\mathbb{E}[g_{JV}(i,S)|R=i,J,B,F=0] - \mathbb{E}[g_{JV}(e,S)|R=i,J,B,F=1] \mid R=i\right\}$
4. $TOT_i = \mathbb{E}\left\{\mathbb{E}[g_{JV}(e,S)|R=i,J,B,F=0] - \mathbb{E}[g_{JV}(e,S)|R=i,J,B,F=1] \mid R=i\right\}$

Intuition for Theorem 1: To provide some intuition for Theorem 1, suppose there is just one analyst and randomization block. We proceed in two steps. First, we show how to decompose total bias into a belief-based and a taste-based term. Second, we show how to further decompose the belief-based term into accurate statistical bias and incorrect belief bias.

A simple difference in mean priority scores across treatment arms among the cases whose original names were European identifies, by ERM and RAND:

$$\begin{aligned} &\mathbb{E}[S|R=e,F=1] - \mathbb{E}[S|R=e,F=0] = \mathbb{E}[\phi_{JiV}(h(i,X))|R=e] - \mathbb{E}[\phi_{JeV}(h(e,X))|R=e] \\ &= \mathbb{E}[\underbrace{\phi_{JeV}(h(i,X)) - \phi_{JeV}(h(e,X))}_{\text{belief-based term}}|R=e] + \mathbb{E}[\underbrace{\phi_{JiV}(h(i,X)) - \phi_{JeV}(h(i,X))}_{\text{taste-based term}}|R=e] \end{aligned}$$

While the first term depends only on belief-based bias $h(e,x) \neq h(i,x)$, the second term depends only on taste-based bias $\phi_{jev}(\cdot) \neq \phi_{jiv}(\cdot)$. A similar decomposition can be applied to the originally-Indigenous cases.

This shows how we can decompose the difference in priority scores into its two constituent parts, but this measures bias in the arbitrary units of the priority score. However, applying any common transformation to S in both arms of the experiment similarly measures total discrimination. In particular, applying $g_{JV}(e,\cdot)$ to both yields:

$$\begin{aligned}
TOT_e &= \mathbb{E}[g_{JV}(e,S)|R=e,F=1] - \mathbb{E}[g_{JV}(e,S)|R=e,F=0] \\
&= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X)))|R=e] - \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h(e,X)))|R=e] \\
&= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X))) - h(e,X)|R=e] \\
&= \mathbb{E}[\underbrace{h(i,X) - h(e,X)}_{\text{belief-based bias}}|R=e] + \mathbb{E}[\underbrace{\phi_{JeV}^{-1}(\phi_{JiV}(h(i,X))) - h(i,X)}_{\text{taste-based bias}}|R=e]
\end{aligned}$$

where we have used Lemma 1 and RAND. Compared with the un-transformed difference in means, $TOT_e := \mathbb{E}[\delta_{JeX}|R=e]$ expresses total discrimination in units of indictments, rather than the arbitrary scale of priority score question. This decomposition also holds for $TOT_i := \mathbb{E}[\delta_{JeX}|R=i]$, the analagous quantity among originally-Inidgenous cases. We focus on the measure δ_{jex} of belief-equivalent bias from Section 3.2 rather than the symmetric form TOT_{jx} solely for simplicity.

The experimental variation allows us to isolate the belief-based component of TOT_e by using *apparent-race* conditional expectations of beliefs $\hat{g}_{jv}(r,s) = \mathbb{E}\left[H|\tilde{R}=r,S=s,J=j,V=v\right]$ rather than applying fixed-race functions $g_{JV}(e,\cdot)$ to priority scores:

$$\begin{aligned}
BBB_e &= \mathbb{E}[g_{JV}(i,S)|R=e,F=1] - \mathbb{E}[g_{JV}(e,S)|R=e,F=0] \\
&= \mathbb{E}[\phi_{JiV}^{-1}(\phi_{JiV}(h_J(i,X)))|R=e] - \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h_J(e,X)))|R=e] \\
&= \mathbb{E}[\underbrace{h_J(i,X) - h_J(e,X)}_{\text{belief-based bias}}|R=e] \qquad (10)
\end{aligned}$$

where we have again used Lemma 1 and RAND. TBB is thus identified by the difference between TOT_e and BBB_e:

$$\begin{aligned}
TBB_e &:= \mathbb{E}[\underbrace{\phi_{JeV}^{-1}(\phi_{JiV}(h_J(i,X))) - h_J(i,X)}_{\text{taste-based bias}}|R=e] = TOT_e - BBB_e \\
&= \{\mathbb{E}[g_{JV}(e,S)|R=e,F=1] - \cancel{\mathbb{E}[g_{JV}(e,S)|R=e,F=0]}\} \\
&\qquad - \{\mathbb{E}[g_{JV}(i,S)|R=e,F=1] - \cancel{\mathbb{E}[g_{JV}(e,S)|R=e,F=0]}\} \\
&= \mathbb{E}[g_{JV}(e,S)|R=e,F=1] - \mathbb{E}[g_{JV}(i,S)|R=e,F=1] \qquad (11)
\end{aligned}$$

The same measures can be computed analogously for originally-Indigenous cases, and then the two can be averaged together to get overall measures TOT, BBB, and TBB.

The second step of the decomposition decomposes belief-based bias BBB_e into the compo-

nent due to accurate statistical bias ASD_e and the part driven by incorrect beliefs IBB_e. To do this, we exploit the fact that equation (10) above shows how to identify the total amount of belief-based bias, while accurate statistical bias ASD_e is identified given our historical administrative data (Y, R, X), allowing us to estimate the function $h(r, x)$ through a machine learning procedure described in detail in Appendix D. This allows us to identify and estimate Incorrect Belief Bias IBB_e as

$$\begin{aligned} IBB_e &:= \mathbb{E}\left[h_J\left(i, X\right) - h\left(i, X\right)\right] - \mathbb{E}\left[h_J\left(e, X\right) - h\left(e, X\right)\right] = BBB_e - ASD_e \\ &= \mathbb{E}\left[g_{JV}\left(i, S\right)|R = e, F = 1\right] - \mathbb{E}\left[g_{JV}\left(i, S\right)|R = e, F = 0\right] \\ &\quad - \mathbb{E}\big\{\mathbb{E}\left[Y|R = i, X\right] - \mathbb{E}\left[Y|R = e, X\right]|R = e\big\} \end{aligned} \tag{12}$$

4.3 A testable implication of the Extended Roy Model

As discussed in section 4.2, the most substantive assumption required for our decomposition is assumption ERM. Assumption ERM has an observable and testable implication that we are able to test in our data. If a given analyst j were to give the same priority score to two cases that share a race and non-race characteristics v (but differ in other non-race characteristics not included in v), then assumption ERM implies that they must also share the same subjective probability of indictment to that analyst. In particular, if x and x' share the same value of v, then:

$$s_j(r, x) = s_j(r, x') \implies h_j(r, x) = h_j(r, x') \tag{13}$$

Equation (13) is implied immediately by Assumption ERM given that the function ϕ_{jvr} is invertible. In fact, equation (13) is essentially equivalent to Assumption ERM: it says that for a given j, r and v, $h_j(r, x)$ can be written as a one-to-one function of $s_j(r, x)$. If one assumes this relationship to be continuous and weakly increasing, this implies that it is in fact strictly increasing (see e.g. Theorem 18.6 of Ross (2003)). Note however that Eq. (13) is not possible in conventional MOT settings in which only a binary action is taken by the agent: invertibility of actions into indictment probabilities is an advantage of our setting in which both can take values on a continuum.

Note that the converse of (13), that for any x and x' that share the same value of v:

$$h_j(r, x) = h_j(r, x') \implies s_j(r, x) = s_j(r, x'), \tag{14}$$

is equivalent to the statement that $s_j(r, x) = \phi_{jrv}(h_j(r, x))$ for *some* function ϕ_{jrv} (not necessarily invertible). This weaker implication is natural in settings with a discrete rather than continuous action space, for example the binary decision to release a defendent on bail. Canay et al. (2023) show that if one assumes that (14) holds for *any* vectors of non-race characteristics x and x' (implying that ϕ_{jr} does not depend on v), the conventional marginal outcomes-test can be used to draw inferences about taste-based discrimination.

This is a strong assumption, which we relax in assumption ERM by allowing ϕ_{jrv} to depend on v. However, this leaves unresolved the question of what v is: which case observables does one need to hold fixed such that priority scores then only rely on the remaining components of x via indictment beliefs $h_j(r, x)$? To allow the data to speak to this question, we observe that assumption ERM carries a testable implication (in settings with a continuous outcome space and beliefs H that are observed up to classical measurement error). In particular, Assumption ERM implies a restriction on observable quantities through equation (13), which can be used to test the assumption for a given choice of v.

To begin, we state a general version of this result which may also be of use in observational settings. Then we specialize to the setting of our lab-experiment for a version of the result that is more easily implemented. Let W be any observable random variable/vector that takes on at least two values w, w' with positive probability, conditional on any values of J, V and $\tilde{R}$.

Proposition 2. *Suppose that assumption ERM holds and W satisfies the following conditions:*

1. *There exist $w, w' \in supp\{W\}$ such that $w, w' \in supp\{W|S, J = j, V = v, \tilde{R} = r\}$ with probability one.*
2. $\mathbb{E}[U|\tilde{R}, J, X, W] = 0$

Then:

$$\mathbb{E}[H|\tilde{R} = r, J = j, V = v, W = w] \\ = \mathbb{E}\left\{\mathbb{E}[H|\tilde{R} = r, S, J = j, V = v, W = w']|\tilde{R} = r, J = j, V = v, W = w\right\} \quad (15)$$

Note that for $w' = w$, equation (15) is trivially satisfied without ERM by the law of iterated expectations. Where ERM comes in is establishing the non-trivial case of equation (15) for $w' \neq w$. In typical non-experimental settings in which $R = \tilde{R}$ and only (J, R, X) is observed by the econometrician (possibly for only a subset of the X), finding a candidate W does require V to be a strict subset of X, so that there can be observable variation in W conditional on $V, \tilde{R}$ and J.

In the case of our triage experiment, we can let $W = R$, and the variation in W given $\tilde{R}$ comes from F, the indicator for a flipped case (e.g. $P(R = e|\tilde{R} = i) = \frac{P(F=1,R=e)}{P(F=0,R=i)+P(F=1,R=e)}$). Note that the second condition $\mathbb{E}[U|\tilde{R}, J, X, W] = 0$ of Proposition 2 is then equivalent to Assumption CLASSICAL. This leads to the following test for ERM in the triage experiment setting:

Corollary 1. *If assumptions ERM and CLASSICAL hold, then for any v and j:*

$$\mathbb{E}[H|R = e, J = j, V = v, F = 1] \\ = \mathbb{E}\left\{ \mathbb{E}[H|R = i, S, J = j, V = v, F = 0] \,\middle|\, R = e, J = j, V = v, F = 1\right\}$$

and

$$\mathbb{E}[H|R=i,J=j,V=v,F=1]$$
$$=\mathbb{E}\left\{\ \mathbb{E}[H|R=e,S,J=j,V=v,F=0]\ \middle|\ R=i,J=j,V=v,F=1\right\}$$

Combining, we have that

$$\mathbb{E}\left\{\ H-\mathbb{E}[H|R=r',S,J,V,F=0]\ \middle|\ R=r,J,V,F=1\right\}=0$$

for each $r\in\{i,e\}$, where r' denotes the opposite value to r.

Note that without measurement error U in subjective indictment probabilities, Assumption ERM would imply that H is degenerate conditional on $\tilde{R},S,J,V$ and thus with probability one $H=\mathbb{E}[H|R,S,J,V,F]$. Among cases with $F=0$, this would imply that $H=\mathbb{E}[H|R,S,J,V,F=0]$ (using RAND). However given that U is mean-zero, a version of this relationship still needs to hold between H and P *on average*.

To implement this test for a given choice of V, we compare elicited beliefs about flipped European cases to the beliefs predicted from unflipped Indigenous cases sharing the same observables V. That is, following corollary 1, we test

$$\mathbb{E}[H-g_{jv}(i,S)|R=e,J=j,V=v,F=1]=0 \tag{16}$$

where $g_{jv}(r,s)$ is defined in equation (8), and similarly for Indigenous cases.[21]

We test the restriction (16) with $v=\emptyset$, by estimating the function $g_j(r,\cdot)=\mathbb{E}[H|R=r,S=\cdot,J=j,F=0]$ separately by analyst j and race r, among the cases seen by each of the analysts in the triage experiment. There are 180 such analysts, each of whom observes around 100 cases in the experiment. For ease of implementation, we assume that $g_j(R,S):=\mathbb{E}[H|R,j,S,F=0]$ is linear in S and its square, with all coefficients allowed to vary by R. We perform inference via a nonparametric bootstrap of this procedure with resampling clustered by analyst. We call an analyst *ERM-friendly* if we fail to reject implication (16) at the 95% level for *both* Indigenous and European cases.

Overall, we reject implication (16) for at least one of $r\in\{i,e\}$ for about half of the 183 analysts. Note that arejection of approximately 10% of analysts would be in line with what we would expect if all analysts satisfied ERM (with $v=\emptyset$), given that we are performing two tests for each analyst with 5% size. However, we still have some reason to suspect that some analysts violate ERM. In Appendix B.2, we test an average of (16) across analysts j, for various choices of v among the observable case variables x, using analyst fixed effects. In this test we

[21] Note that if we considered unflipped, rather than flipped European cases, the analogous equation

$$\mathbb{E}[H|R=e,J=j,V=v,F=0]=\mathbb{E}[g_{jv}(i,S)|R=e,J=j,V=v,F=0]$$

would hold immediately by the law of iterated expectations, without requiring Assumption ERM.

find significant violations of the averaged implication of ERM whether v is empty or includes indicators for case typology, geography, or both. This suggests that ERM violations do occur among the analysts in our sample, but these average violations may be driven by relatively few such analysts. For this reason, we present our main results both for all analysts, and for the 143 *ERM-friendly* analysts.[22]

A limitation of the ERM test is that it could reject even if ERM holds, provided that other maintained assumptions fail. For example, we could reject due to non-classical measurement error in observing beliefs, or due to discreteness of S. For this reason, we present results both with and without restricting the analysis to ERM-friendly analysts, for comparison.

4.4 "As-if" interpretations of the ERM restriction

Assumption ERM takes a stand on what information analysts use in generating their priority scores S in the lab-experiment: the information denoted X displayed for the denuncia. Specifically, ERM is an assumption about the structure of the function $s_j(r, x)$, which describes counterfactual priority scores would take for alternative values x of X. In this notation, $S = s_J(\tilde{R}, X)$.

Specifically, Assumption ERM imposes that $s_j(r, x) = \phi_{jrv}(h_j(r, x))$ for a strictly increasing function ϕ_{jrv}. Given ERM, we define taste-based bias as any dependence of the function ϕ on race r (in line with Canay et al. (2023) for the case of a binary action).

Focusing as well on binary actions, Hull (2021) considers an alternative approach to defining taste-based bias (termed therein as *canonical taste-based discrimination*) that remains agnostic about what information set the decision-maker uses to choose their action. Under this approach, a decision-maker exhibits TBB if there exists an information set such that their actions are consistent with taste-based choice under that information set. This approach also allows the marginal outcome test to be informative about TBB without an ERM restriction.

Such an approach is less appropriate in our setting where we have precise control over the information set provided to the analysts in our lab experiment. However, given the restrictiveness of the ERM assumption, it is interesting to consider whether the approach of Hull (2021) can be used to relax Assumption ERM in our setting.

In particular, Proposition 3 of Hull (2021) shows that given a decision rule of the general form $\mathbb{1}[W \leq 0]$, decisions can be rationalized by a ranking W_i of the form $W = \mathbb{E}[Y^*|\tilde{R}, V] - t(\tilde{R})$ for *some* "signal" V, if and only if $\mathbb{E}[Y^*|\tilde{R} = r, W = u]$ is strictly increasing in u for each r (where Y^* denotes potential for pre-trial misconduct).[23] Notably, this form for W yields a decision rule $\mathbb{1}[\mathbb{E}[Y_i^*|\tilde{R}, V] \leq t(\tilde{R})]$ that is of the extended Roy Model form discussed by Canay et al. (2023). Thus, one can say that if the marginal treatment effect curves $\mathbb{E}[Y^*|R = r, U = u]$ are strictly

[22] Appendix B.3 shows that in principle one can implement a systematic search for subsets v of the observables that are sufficient to make Assumption ERM hold. Since v does not seem to be required to pass the analyst-specific test for the majority of analysts, we do not implement this search.

[23] I use the notation W here to avoid confusion with our measurement error U in belief elicitation. Hull (2021) instead uses a decision-rule with the notation $\mathbb{1}[U \leq 0]$ and then defines $W = F_{U|R}(U)$. Note that with this definition $\mathbb{E}[Y^*|\tilde{R} = r, W]$ is strictly increasing in u (for a given r) iff $\mathbb{E}[Y^*|\tilde{R}, U = u]$ is strictly increasing in u.

increasing in u, there exists a V such that ERM holds with $X = V$.

A similar result can be obtained in our continuous-action framework, that is in fact more readily applicable since our analog of the MTE curve, the function g, can be identified globally and without using instrumental variables.[24] Consider the case with $v = \emptyset$ for simplicity, and suppose that $g_j(r, s) = \mathbb{E}[H|\tilde{R} = r, S = s, J = j, F = 0]$ is strictly increasing in s for each r and j. Define ϕ_{jr} to be the inverse of $g_j(r, \cdot)$, which is then guaranteed to exist. It follows then that $S = \phi_{J\tilde{R}}(\mathbb{E}[H|\tilde{R}, S, J, F = 0])$ with probability one. Substituting in $S = s_J(\tilde{R}, X)$, we have then that $S = \phi_{J\tilde{R}}(\mathbb{E}[H|\tilde{R}, s_J(\tilde{R}, X), J, F = 0])$, i.e. agents act *as if* Assumption ERM holds with the potential outcomes function $h_j(r, x) = \mathbb{E}[H|r, s_j(r, x), j, F = 0]$ for beliefs.

However, since we additionally observe beliefs $H^* = h_J(\tilde{R}, X)$ up to the measurement error U where $H = H^* + U$, we can ask the question of whether this interpretation would be consistent with our observed belief data. That is, given Assumption CLASSICAL, is it the case that

$$\mathbb{E}[H|\tilde{R}, X, J, F] = \mathbb{E}\left\{\mathbb{E}[H|\tilde{R}, S, J, F = 0]|\tilde{R}, X, J, F\right\}$$

This equation is trivially satisfied for $F = 0$ by the law of iterated expectations, since by assumption S is deterministic conditional on $\tilde{R}$, X, and J. Alternatively, for $F = 1$, the above consistency condition is

$$\mathbb{E}[H|\tilde{R}, X, J, F = 1] = \mathbb{E}\left\{\mathbb{E}[H|\tilde{R}, S, J, F = 0]|\tilde{R}, X, J, F = 1\right\}$$

which is exactly the set of testable restrictions in Corollary 1.

4.5 The value of data on beliefs in decomposing bias

Our decomposition result in theorem 1 relies on observing H, analysts' subjective indictment probabilities, in order to estimate the function g. In previous settings studying disparate treatment, researchers have generally lacked data on beliefs and have used a downstream outcome available from observational administrative data to implement the marginal outcomes test. However, this strategy relies on the distribution of the outcome available to the econometrician appropriately matching the probabilities that agents use to drive their actions.

In our setting a natural downstream outcome is the indictment Y. Accordingly, define $p_{jv}(r, s) := \mathbb{E}[Y|R = r, S = s, J = j, V = v, F = 0]$. To see where the results of theorem 1 may fail when replacing the conditional expectation of beliefs g with the conditional expectation of indictment p, consider first the task of identifying taste-based bias through (9), We have the following analog of Lemma 1:

[24] g plays the analog of $\mu(r, u) := \mathbb{E}[Y^*|\tilde{R} = r, W = u]$ in the sense that if beliefs are accurate about Y^*, then within cases of apparent race r (and fixing a single analyst), both $\mu(r, W_i)$ and $g(r, S_i)$ rank cases in order of $\mathbb{E}[Y_i|R_i = r, X_i = x]$ if the analyst's information set is X_i.

Lemma 2. *Under ERM:*

$$p_{jv}(r,s) = \phi^{-1}_{jrv}(s) - \mathbb{E}[h_j(r,x) - h(r,x)|R=r, S=s, J=j, V=v, F=0]$$

Thus $p_{jv}(e,s) - p_{jv}(i,s)$ may differ from $\phi^{-1}_{jev}(s) - \phi^{-1}_{jiv}(s)$ unless

$$\mathbb{E}[\tilde{h}_j(i,X)|R=i, S=s, J=j, V=v, F=0] = \mathbb{E}[\tilde{h}_j(e,X)|R=e, S=s, J=j, V=v, F=0]$$

where $\tilde{h}_j(r,x) := h_j(r,x) - h(r,x)$. If analysts have correct beliefs about indictment probabilities, in accordance with the population from which the data is drawn, then $\tilde{h}_j(r,x) = 0$ and the above equation will be satisfied.

This has two implications. First, without data on analysts' beliefs, the econometrician is only able to estimate the extent of taste-based bias and distinguish it from belief-based bias when analysts hold correct beliefs. Conversely, when analysts do not hold correct beliefs (as we will show that they do not), with data on beliefs, the econometrician is able to correctly distinguish taste-based and belief-based bias, and further is able to distinguished accurate statistical bias from incorrect belief bias (Theorem 1).

Second, we can show how to use only data on outcomes to implement the decomposition into taste-based and belief-based bias, and we can show how to sign and quantify the bias that results whenever analysts hold incorrect beliefs. We state the following result:

Theorem 2. *Under ERM and RAND, the estimands analagous to those in Theorem 1 but with Y replacing H satisfy:*

$$\mathbb{E}\left\{\mathbb{E}[p_{JV}(i,S)|R=e,J,B,F=1] - \mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=0]|R=e\right\} = BBB_e + W^{\mathcal{B}}_e$$
$$\mathbb{E}\left\{\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=1] - \mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=0]|R=e\right\} = TOT_e + W^{T}_e$$
$$\mathbb{E}\left\{\mathbb{E}[p_{JV}(i,S)|R=i,J,B,F=0] - \mathbb{E}[p_{JV}(e,S)|R=i,J,B,F=1]|R=i\right\} = BBB_i + W^{\mathcal{B}}_i$$
$$\mathbb{E}\left\{\mathbb{E}[p_{JV}(e,S)|R=i,J,B,F=0] - \mathbb{E}[p_{JV}(e,S)|R=i,J,B,F=1]|R=i\right\} = TOT_i + W^{T}_i$$

where the bias terms $W^{\mathcal{B}}_e = W^{\mathcal{B}}_i = W^{T}_e = W^{T}_i = 0$ if $h_j(r,x) = h(r,x)$ for all j,r,x (accurate beliefs).

The bias terms $W^{\mathcal{B}}_r$ and W^{T}_r arise from misattributing analysts' systematically incorrect beliefs to belief-based bias by assuming that their beliefs are correct. These bias terms may be small if the distribution of scores among originally European and originally Indigenous cases are not too different. But if lab-analysts tend to over-estimate the probability of indictment more for Indigenous cases than they do for European cases, for example, then $\tilde{h}_j(i,x) > \tilde{h}_j(e,x)$. The expressions for $W^{\mathcal{B}}_r$ and W^{T}_r in the proof of Theorem 2 show that our estimate of TBB will then be biased upwards, which may misleadingly suggest taste-based discrimination against European defendants even if taste-based bias exists against Indigenous defendants.

5 Results

5.1 Reduced form results from the triage experiment

Table 1 reports our results from the experiment outlined in Appendix C. We estimate the impact of experimentally manipulating defendants' surnames on analysts' assigned scores with report-level and analyst-level controls:

$$\begin{aligned}\text{score}_i = \alpha + \beta_1 \text{Switch Euro to Ind}_i + \beta_2 \text{Switch Ind to Euro}_i + \beta_3 \text{Original Ind}_i \\ + \mathbf{X_i} + \mathbf{X_j} + \varepsilon_i\end{aligned}$$

where score_i is the score assigned to report i, $\text{Switch Euro to Ind}_i$ and $\text{Switch Ind to Euro}_i$ indicate whether the surname was experimentally switched from European to Indigenous or from Indigenous to European, respectively, and Original Ind_i indicates the defendant's original surname was Indigenous. $\mathbf{X_i}$ and $\mathbf{X_j}$ are vectors of report and analyst controls.[25]

Our results show that on average, analysts who see a complaint presented as accusing an Indigenous defendant assign it a higher indictment probability score than those who see it as a European defendant.

	(1)	(2)	(3)	(4)
Switching surname from European to Indigenous	1.544***	1.255***	1.063**	0.835**
	(0.484)	(0.416)	(0.461)	(0.390)
Switching surname from Indigenous to European	-1.223	-0.346	-0.282	0.584
	(0.907)	(0.807)	(0.855)	(0.742)
Original Indigenous surname	-4.225***	-4.647***		
	(0.574)	(0.503)		
Analyst-level controls	Yes	No	Yes	No
Complaint-level controls	Yes	Yes	No	No
Analysts FEs	No	Yes	No	Yes
Complaints FEs	No	No	Yes	Yes
Clusters (analyst-level SEs)	177	182	177	182
Adjusted R-squared	0.029	0.427	0.130	0.530
Observations	18,706	19,112	18,706	19,112

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 1: RESULTS OF LAB-IN-THE-FIELD TRIAGE EXPERIMENT. OUR DEPENDENT VARIABLE IS INDICTMENT PROBABILITY SCORE, WITH MEAN=63.01 AND S.D.=28.96.

[25] We can modify our specification to include report and analyst fixed effects:

$$\text{Score}_i = \alpha + \beta_1 \text{Switch_Euro_to_Ind}_i + \beta_2 \text{Switch_Ind_to_Euro}_i + \phi_i + \phi_j + \varepsilon_i$$

where ϕ_i are report fixed effects and ϕ_j are analyst fixed effects. This approach controls for all time-invariant characteristics at both the report and analyst levels.

5.2 Results of the decomposition into BBB and TBB

Figure 2 and Table 2 report our main results for belief based and taste-based bias, based on the identification results established in Section 4. The second column of table 2 restricts the analysis to "ERM-friendly" analysts that pass the overidentification restriction described in Section 4.3. The estimates suggest that on average, analysts exhibit taste-based bias against Indigenous defendants along with belief-based bias in favor of Indigenous defendants, leading to an overall total bias weakly against Indigenous defendants. The patterns hold whether averaging over originally-European or originally-Indigenous cases. Table 3 shows a similar pattern holds among Indigenous and European analysts.

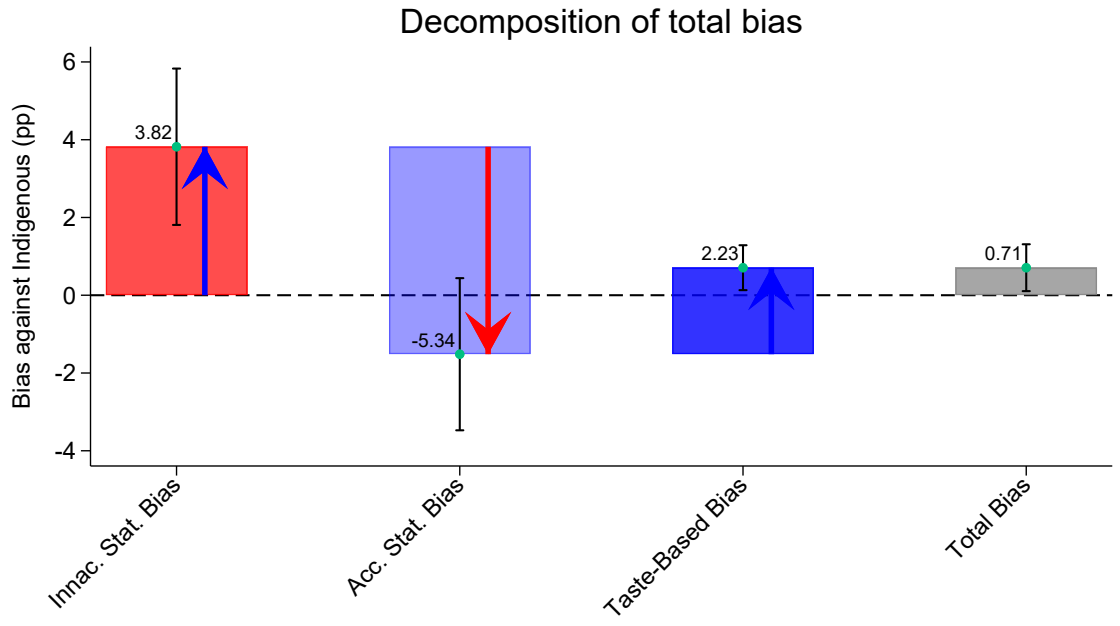


FIGURE 2: MAIN RESULT FOR THE DECOMPOSITION OF TOTAL BIAS. BIAS QUANTITIES ARE IN UNITS OF PERCENTAGE POINTS OF INDICTMENT PROBABILITY.

5.3 Results on inaccuracy in analysts' beliefs

Table 4 shows estimates of $\mathbb{E}[H - \hat{h}(R, X)|R = r, J \in \mathcal{J}]$ for given races $r \in \{i, e\}$ of the original historical case, and sets $\mathcal{J}$ of analysts (Indigenous analysts, European analysts, and all analysts). Given Assumption CLASSICAL, and assuming consistency of the random forest predictions, $\mathbb{E}_n[H - \hat{h}(R, X)|R = r, J \in \mathcal{J}]$ is a consistent estimator for $\mathbb{E}[\tilde{h}_j(r, x)|R = i, J \in \mathcal{J}] := \mathbb{E}[h_J(r, x) - h(r, x)|R = i, J \in \mathcal{J}]$, where $\mathbb{E}_n$ is the sample expectation function.

In particular, for a given choice of $\mathcal{J}$, the second row of Table 4 shows our estimate of $\mathbb{E}[\tilde{h}_j(r, x)|R = i, J \in \mathcal{J}]$ (average inaccuracy in beliefs among Indigenous cases). The second row displays estimates of $\mathbb{E}[\tilde{h}_j(r, x)|R = e, J \in \mathcal{J}] - \mathbb{E}[\tilde{h}_j(r, x)|R = i, J \in \mathcal{J}]$. The results indicate that analysts *overestimate* indictment probabilities (by about 0.12 or 12 points for Indigenous

	(1) All analysts	(2) ERM-friendly analysts
Total bias	0.708*	0.330
	(0.330)	(0.400)
Taste-based bias	2.225***	1.601***
	(0.325)	(0.381)
Belief-based bias (total)	-1.517***	-1.271*
	(0.435)	(0.523)
Accurate statistical bias	-5.336***	-5.534***
	(0.856)	(1.173)
Inaccurate statistical bias	3.819***	4.264***
	(0.919)	(1.264)
Error term (total bias)	-0.541*	-0.246
	(0.256)	(0.297)
Error term (taste-based bias)	-4.423***	-4.296***
	(0.842)	(1.031)
Error term (belief-based bias)	3.882***	4.050***
	(0.887)	(1.091)
Restricted to ERM-friendly analysts		X
Number analysts	182	95
Number of obs.	19112	11346

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 2: RESULTS OF THE BIAS DECOMPOSITION.

cases, and by roughly 0.08 or 8 points for European cases). The gap of roughtly 0.1 (or 10 points on a 100 point scale) is strongly statistically in all three columns. Thus the inaccurate statistical distrimiination quantity IBB (labelled "differentially-incorrect indictment beliefs" in Eq. (I.2)) is *positive*, meaning that belief-based bias $h_j(i,x) - h_j(e,x)$ tends to be higher (i.e. more biased against Indigenous) than accurate statistical discrimination $h(i,x) - h(e,x)$ (which turns out to be negative, i.e. biased against European).

Incorrectly assuming that analysts' beliefs are accurate and using Y in the decomposition of total bias into BBB and TBB, would attribute this gap to BBB, yielding overestimates of BBB and underestimates of TBB. In fact, since the error term for taste based bias is negative and larger than the positive estimate of true TBB, we see from Table 2 that using Y instead of H would result in estimates of TBB of exactly the wrong sign: suggesting TBB againts Europeans rather than against Indigenous.

	(1) European analysts	(2) Indigenous analysts
Total bias	0.138	0.925
	(0.505)	(0.780)
Taste-based bias	1.322**	2.480**
	(0.427)	(0.837)
Belief-based bias (total)	-1.184	-1.555
	(0.683)	(1.178)
Accurate statistical bias	-5.025***	-7.168**
	(1.436)	(2.603)
Inaccurate statistical bias	3.841**	5.614*
	(1.367)	(2.833)
Error term (total bias)	-0.101	-0.722
	(0.372)	(0.630)
Error term (taste-based bias)	-4.482***	-3.590
	(1.254)	(2.214)
Error term (belief-based bias)	4.381***	2.867
	(1.308)	(2.309)
Restricted to ERM-friendly analysts	X	X
Number analysts	73	22
Number of obs.	8667	2679

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 3: DECOMPOSITION RESULTS BY RACE OF HISTORICAL ANALYST . RELATIVE TO TABLE 2, THIS TABLE IS SIMPLIFIED BY ONLY CONSIDERING ESTIMATES BASED ON STATED BELIEFS H, ONLY ESTIMATING ON THE RESTRICTED SAMPLE OF ERM-FRIENDLY ANALYSTS, AND AVERAGING BIAS ACROSS ORIGINALLY-EUROPEAN AND ORIGINALLY-INDIGENOUS CASES.

	(1) All analysts	(2) European analysts	(3) Indigenous analysts
Difference Euro.-Indig. cases	-3.819***	-3.802**	-3.880
	(1.061)	(1.208)	(2.212)
Indigenous cases	11.88***	11.14***	14.19***
	(0.948)	(1.078)	(1.990)
N	14506	10971	3535

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 4: ESTIMATES OF THE AVERAGE BIAS IN BELIEFS ABOUT THE INDICTMENT PROBABILITY FOR CASES OF A GIVEN ETHNICITY, MEASURED OUT OF 100 POINTS. SEE TEXT ABOVE FOR DESCRIPTION.

6 Policy exercise

The decomposition in Section 5 estimates each analyst's bias at the individual level. For a principal who observes these estimates, a natural policy exercise is the assignment of analysts to cases. Such reassignment is potentially valuable because the heterogeneity we document is substantial: some analysts exhibit pronounced taste-based bias against Indigenous defendants, others exhibit little or none, and a few are biased in the opposite direction.[26] Assigning cases to less biased analysts can reduce the differential treatment that Indigenous defendants experience in practice. However, reassignment on equity grounds may divert cases from analysts who are more productive at detecting corruption. Our estimates allow us to quantify this bias-productivity trade-off by combining analyst-specific measures of bias with analyst-case measures of productivity.

We formalize our analyst–case matching as a linear assignment that penalizes group–specific bias while steering assignments toward higher downstream productivity under a fixed escalation rule. Analysts are indexed by $j \in \{1, \ldots, J\}$ and cases by $c \in \{1, \ldots, n\}$. Each analyst may handle up to k cases. For analyst–case pair (j, c) we observe a signed bias C_{jc} such that positive means bias against Indigenous cases, and negative means bias against European.[27] An analyst priority score is denoted by S_{jc}. For each case c we also have an expected indictment probability $p_c \in [0, 1]$.[28] We decompose bias into one–sided, nonnegative magnitudes

$$I_{jc} = \max\{C_{jc}, 0\}, \qquad E_{jc} = \max\{-C_{jc}, 0\},$$

so that I_{jc} measures bias against Indigenous and E_{jc} measures bias against European.

[26] If bias were uniform across analysts, reassignment would offer no scope for improvement.

[27] The choice of C_{jc} depends on which component of the decomposition the principal seeks to minimize. Setting $C_{jc} = \mathrm{TBB}_{jc}$ targets taste-based bias alone; $Cjc = \mathrm{TBB}_{jc} + \mathrm{IBB}_{jc}$ additionally penalizes incorrect beliefs; and $C_{jc} = \mathrm{TOT}_{jc}$ penalizes all forms of differential treatment. Each choice produces a different optimal assignment and a different equity–productivity frontier.

[28] Estimated from the classification model in Appendix D.

An assignment is a binary matrix $A = [a_{jc}]$ with $a_{jc} \in \{0, 1\}$ and $\sum_j a_{jc} = 1$ for each real case c, enforcing that every case is assigned to exactly one analyst without splitting. Capacity is handled by slot expansion: replicating each analyst into k slots to obtain $N = nk$ rows, adding $N - n_a$ dummy case columns (zero cost) to square the matrix, and imposing $\sum_c a_{\sigma c} = 1$ for each slot σ and $\sum_\sigma a_{\sigma c} = 1$ for each (real or dummy) case. This puts the problem into the standard assignment form that can be solved by the Hungarian algorithm.

After solving an assignment, the policy *escalates* the N_{top} cases with the largest match–specific scores

$$s_c(A) \;=\; S_{j(c),\,c} \;=\; \sum_{j=1}^{n_a} S_{jc}\, a_{jc},$$

where $j(c)$ is the analyst assigned to case c. Denote the indices of these N_{top} cases by $\mathcal{T}_{N_{\text{top}}}(A)$.[29] Productivity is then evaluated as the expected indictments of the escalated set,

$$\text{Output}(A) \;=\; \sum_{c \in \mathcal{T}_{N_{\text{top}}}(A)} p_c,$$

while bias on the escalated set is summarized by

$$\overline{I}(A) \;=\; \frac{1}{N_{\text{top}}} \sum_{c \in \mathcal{T}_{N_{\text{top}}}(A)} I_{j(c)c}, \qquad \overline{E}(A) \;=\; \frac{1}{N_{\text{top}}} \sum_{c \in \mathcal{T}_{N_{\text{top}}}(A)} E_{j(c)c}.$$

Our goal is to minimize group–bias while rewarding productivity of the *escalated* set,

$$\min_A \Big(\underbrace{\sum_{j,c} \big(w_I I_{jc} + w_E E_{jc}\big) a_{jc}}_{\text{Bias}(A)} \;-\; w_O\, \text{Output}(A) \Big).$$

The difficulty is that $\text{Output}(A)$ depends on which N_{top} cases clear the assignment–dependent threshold $s_c(A) \geq \tau(A)$. This top–N selection is a non-separable operator, so the cost can no longer be written as a sum of per–pair terms $\sum_{j,c} \text{cost}_{jc}\, a_{jc}$. Placing $\text{Output}(A)$ directly inside the minimand means that methods for solving linear assignment problems like the Hungarian algorithm, which require a separable linear objective in the a_{jc}, cannot be applied directly.

Proxy objective and solution. We cannot simply reward high p_c by subtracting a case–only term $-\sum_{j,c} w_P\, p_c\, a_{jc}$ because in any feasible assignment $\sum_j a_{jc} = 1$ for each real case c, we have

$$\sum_{j,c} p_c a_{jc} = \sum_c p_c \Big(\sum_j a_{jc} \Big) = \sum_c p_c,$$

which is a *constant* independent of A. Changing w_P has no effect on the solution and so we

[29] Note: in case of too many ties we would require a tie breaking rule.

cannot trace the efficiency–bias curve.

We therefore use a pairwise proxy that depends on both the analyst and the case: the interaction $p_c S_{jc}$, which rewards matching high–p_c cases to analysts who would score them highly, increasing their chance of landing in $\mathcal{T}_{N_{\text{top}}}(A)$ after assignment.

We solve the linear assignment

$$\min_{A} \sum_{j,c} \Big(w_I \, I_{jc} \; + \; w_E \, E_{jc} \; - \; w_{PS} \, p_c S_{jc} \Big) \, a_{jc},$$

after the matrix capacity expansion. The bias terms enter positively so minimization reduces them; the pairwise productivity proxy enters negatively so minimization prefers pairs with large $p_c S_{jc}$, i.e., productive cases placed with analysts who give them a higher rank. The Hungarian algorithm delivers an exact optimum for this separable objective. We then apply the policy rule to the realized scores $s_c(A)$, escalate the top N_{top} cases, and compute the true productivity $\text{Output}(A) = \sum_{c \in \mathcal{T}_{N_{\text{top}}}(A)} p_c$ together with $\overline{I}(A)$ and $\overline{E}(A)$ on that escalated set.

Holding w_I and w_E fixed and increasing w_{PS} shifts the assignment toward pairings that are more likely to push high–p_c cases into the escalated set, typically increasing realized $\text{Output}(A)$ at the cost of higher bias; decreasing w_{PS} does the opposite. Because the realized output is evaluated exactly after escalation, the curve we trace reflects the true productivity of the implemented policy, even though the proxy $p_c S_{jc}$ is used inside the assignment to remain compatible with the Hungarian algorithm.

The proxy is an approximation to the nonseparable objective; it performs best when S_{jc} is indeed the operative ranking for escalation and when p_c is informative about productivity. Priority score and probability are not on the same scale: if priorities or probabilities have very different scales we can adjust weights or rescale inputs to avoid numeric dominance.[30] Ties at the N_{top} boundary are resolved by a fixed rule to ensure a well–defined $\mathcal{T}_{N_{\text{top}}}(A)$.

While the pairwise proxy p_c, S_{jc} keeps the objective linear and compatible with the Hungarian algorithm, it remains an approximation. The true productivity depends on the top-N set selected after assignment, not on per-pair terms that sum separably across (j, c). As a result, the proxy can mis-rank trade-offs, and it cannot encode the escalation rule or impose global constraints on the escalated set (such as a bias cap or output floor) inside the optimization; these are only checked ex post. Tuning w_{PS} therefore traces an informative but approximate efficiency–bias frontier.

Hungarian algorithm with an ex-ante threshold. In our original problem, escalation of top-N cases by assignment dependent scores $s_c(A) = S_{j(c)c}$ is a step function at a threshold $\tau(A)$ that depends on assignment. However, the Hungarian algorithm requires a separable objective $\sum_{j,c} \text{cost}_{jc} a_{jc}$. To approximate this step rule for escalation, we fix a threshold $\tau^\star$ ex ante, using

[30] S_{jc} or p_c can be rescaled (e.g., by an affine transformation) so that the weights become more interpretable.

our historical administrative data (e.g., score boundary that on average separates escalated from non-escalated cases). We then reward only those analyst-case pairs whose score would clear this threshold. Consider two options:

Let

$$\begin{aligned} \phi_{\tau^\star}(S) &= \mathbf{1}\{S \geq \tau^\star\} \qquad \text{(hard step)} \qquad \text{or} \\ \phi_{\tau^\star}(S) &= \max\{S - \tau^\star, 0\} \quad \text{(soft ramp).} \end{aligned}$$

Then our linear assignment objective becomes:

$$\min_{A} \sum_{j,c} \Big(w_I \, I_{jc} + w_E \, E_{jc} \; - \; w_\tau \, p_c \, \phi_{\tau^\star}(S_{jc}) \Big) \, a_{jc},$$

which is separable in (j, c) and thus solvable exactly by the Hungarian algorithm after the usual matrix capacity expansion. Intuitively, the model now prefers matches that both (i) have low one-sided bias magnitudes and (ii) are likely to place the case above the threshold that triggers escalation, weighted by the expected indictments conditional on escalation, p_c.

After solving for the assignment we escalate the top N cases by the realized scores $s_c(A) = S_{j(c),c}$ and evaluate productivity as $\mathrm{Output}(A) = \sum_{c \in \mathcal{T}_N(A)} p_c$. The threshold proxy shapes the assignment to mimic the step function, while the realized output is computed without approximation. This better matches the shape of the escalation rule.

A caveat is that a global $\tau^\star$ from administrative data ignores that S_{jc} distributions might differ across analysts, favoring those with systematically higher scores. To mitigate this, we can use analyst-specific thresholds $\tau_j^\star$ derived from their historical distributions, replacing $\phi_{\tau^\star}(S_{jc})$ by $\phi_{\tau_j^\star}(S_{jc})$; the objective remains separable. Note that fixing $\tau^\star$ does not guarantee exactly N_{top} cases above the threshold once the assignment is chosen; we still escalate the top N_{top} cases, but for proxy calibration one should iterate $\tau^\star$ so that the share above it is approximately N_{top}/n_a. Finally, if $\tau^\star$ is set too high or too low, few or almost all pairs clear the threshold, which is why the soft ramp formulation may be more forgiving.

Mixed-integer linear program

The limitations of the proxy motivate a formulation that encodes the escalation rule directly. We introduce binary escalation variables $z_c \in 0, 1$ indicating whether case c is escalated, a continuous score s_c equal to the realized priority of case c under its assigned analyst, and an endogenous threshold τ separating escalated from non-escalated cases. The program jointly optimizes the assignment and the escalation set:

$$\min_{a,z,s,\tau} \sum_{j,c} \big(w_I, I_{jc} + w_E, E_{jc} \big), a_{jc}; -; w_O \sum_{c=1}^{n_c} p_c, z_c,$$

subject to

$$\sum_j a_{jc} = 1 \quad \forall, c, \qquad \sum_c a_{jc} \le k \quad \forall, j,$$

$$s_c = \sum_j S_{jc}, a_{jc} \quad \forall, c,$$

$$\sum_c z_c = N_{\text{top}},$$

$$s_c; \ge; \tau - M(1 - z_c) \quad \forall, c, \qquad s_c; \le; \tau - \varepsilon + M, z_c \quad \forall, c,$$

$$a_{jc} \in 0, 1, \quad z_c \in 0, 1, \quad s_c \in R, \quad \tau \in R.$$

Here M is any constant that upper-bounds the range of S (e.g. $M \ge S_{\max} - S_{\min}$ with known bounds $S_{\min} \le S_{jc} \le S_{\max}$), and $\varepsilon > 0$ is a small tie-break. The two s_c inequalities enforce the exact top-N by S rule: whenever $z_c = 1$ we must have $s_c \ge \tau$, and whenever $z_c = 0$ we must have $s_c \le \tau - \varepsilon$. Combined with $\sum_c z_c = N_{\text{top}}$, the model chooses a threshold τ that selects precisely N_{top} cases with the largest assignment-dependent scores, resolving ties deterministically via ε. Unlike the Hungarian approach, there is no proxy for productivity in the objective: the term $-w_O \sum_c p_c, z_c$ is the exact expected number of indictments in the escalated set.

To measure bias on the escalated set within the optimization, we introduce auxiliary variables $b_{jc} = a_{jc}, z_c$, representing the product of the assignment and escalation indicators. Since both a_{jc} and z_c are binary, we linearize their product via the standard McCormick constraints

$$0 \le b_{jc} \le a_{jc}, \quad 0 \le b_{jc} \le z_c, \quad b_{jc} \ge a_{jc} + z_c - 1 \qquad \forall, j, c.$$

Then $\sum_{j,c} I_{jc}, b_{jc}$ equals the total Indigenous-unfavorable bias magnitude among escalated cases, and $\sum_{j,c} E_{jc}, b_{jc}$ the European-unfavorable magnitude. This permits explicit caps on escalated-set bias,

$$\sum_{j,c} I_{jc}, b_{jc}; \le; \bar{x} I, N_{top}, \qquad \sum_{j,c} E_{jc}, b_{jc}; \le; \bar{x} E, N_{top},$$

or a productivity floor $\sum_c p_c, z_c \ge \underline{Y}$, while still minimizing the primary objective. These global constraints are nonseparable and cannot be represented in a linear assignment; the MILP handles them directly. After solving, the realized escalated-set summaries

$$\overline{I} = \frac{1}{N_{\text{top}}} \sum_{j,c} I_{jc}, b_{jc}, \qquad \overline{E} = \frac{1}{N_{\text{top}}} \sum_{j,c} E_{jc}, b_{jc}, \qquad \text{Output} = \sum_c p_c, z_c$$

are obtained directly from the solution (a, z, b).

The CGR workflow is inherently two-stage: an analyst first escalates a subset of cases to more in-depth investigation, and only escalated cases can subsequently be indicted. The MILP formulation mirrors this structure exactly. The assignment variables a_{jc} determine which ana-

lyst reviews each case, and the escalation variables z_c, constrained to select exactly the top N_{top} cases by the assignment-dependent scores $s_c(A) = \sum_j S_{jc}, a_{jc}$ via the threshold τ and big-M linking constraints, determine which cases advance. Because p_c is estimated as the probability of indictment conditional on escalation and case features (Appendix D), the objective term $-w_O \sum_c p_c, z_c$ computes the expected number of indictments on the escalated set precisely as the CGR process would deliver it. Productivity therefore accrues only from escalated cases, and those cases are escalated by the analyst-score ranking induced by the assignment.

A further advantage is that policy trade-offs are encoded exactly rather than approximately. Because escalation is modeled explicitly, the MILP can accommodate global constraints that the Hungarian approach cannot: a minimum output target, caps on escalated-set bias, analyst workload limits, or quotas by case type. The reported output and group-bias summaries are the quantities the institution would realize under the recommended assignment, with no proxy step. The main cost is computational: the MILP is NP-hard and slower than the Hungarian algorithm's polynomial runtime. In practice, we warm-start the solver at the Hungarian solution, which provides a feasible starting point and substantially reduces solve times.

7 Conclusion

In this paper we have provided a comprehensive framework within which to study biases in how citizens are treated by the state. The framework incorporates the three main sources of bias discussed in the economics literature: Biases based on tastes, on statistical discrimination, and on inaccurate beliefs (Arnold et al., 2018; Arnold et al., 2022; Bohren et al., 2019; Lang and Lehmann, 2012b; Hurst et al., 2024; Bohren et al., 2025). It also incorporates citizens' sequential encounters with state actors and the interlinkages between the biases at each stage. We operationalize the framework using a triage experiment conducted with Peruvian anti-corruption officials, finding that assuming analysts to have accurate beliefs about indictment suggests belief-based bias against Indigenous defendants along with taste-based bias against Europeans, while using elicited beliefs suggests the opposite: belief-based bias against European cases along with taste-based bias against Indigenous defendants, leading to a small total bias against Indigenous defendants.

We reconcile these findings by estimating "true" indictment probabilities from a machine learning prediction of indictment given the rich covariates we observe about each case, defined from the full information displayed to analysts in the lab experiment. Our estimates suggest that analysts systematically overestimate the probability of indictment for cases, but they do so more for Indigenous defendants than for European defendants. This implies that belief-based bias against Indigenous cases is smaller than the "accurate statistical discrimination" that analysts would employ if they had accurate beliefs. We estimate that accurate statistical discrimination would in fact be against European defendants. Incorrectly assuming accurate beliefs attributes this differential to belief based bias, leading to misleading estimates for both belief-based and

taste-based bias. This underscores the value of observing beliefs when implementing marginal outcomes tests.

References

AIGNER, D. J. and CAIN, G. G. (1977). "Statistical Theories of Discrimination in Labor Markets". *ILR Review* 30, pp. 175–187.

ARNOLD, D., DOBBIE, W. and HULL, P. (2022). "Measuring Racial Discrimination in Bail Decisions". *American Economic Review* 112 (9), pp. 2992–3038.

ARNOLD, D., DOBBIE, W. and YANG, C. S. (May 2018). "Racial Bias in Bail Decisions". *The Quarterly Journal of Economics* 133 (4), pp. 1885–1932.

ARROW, K. J. (1973). "The Theory of Discrimination". *Discrimination in Labor Markets*. Ed. by O. ASHENFELTER and A. REES. Princeton University Press.

ARTILES, M. (2024). "Ethnic Diversity, Historical Economic Exchange, and Development: Evidence from Andean Peru". Working paper.

BANK, W. (2023). *World Bank Enterprise Surveys: Peru*. Development Data Group - World Bank Group.

BARON, E. J., DOYLE JR, J. J., EMANUEL, N., HULL, P. and RYAN, J. (2024). "Discrimination in multiphase systems: Evidence from child protection". *The Quarterly Journal of Economics* 139 (3), pp. 1611–1664.

BECKER, G. S. (1971). *The economics of discrimination*. University of Chicago Press.

BERTRAND, M. and DUFLO, E. (2017). "Field experiments on discrimination". *Handbook of Economic Field Experiments* 1, pp. 309–393.

BERTRAND, M. and MULLAINATHAN, S. (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". *The American Economic Review* 94 (4), pp. 991–1013.

BOHREN, J. A., HAGGAG, K., IMAS, A. and POPE, D. G. (2025). "Inaccurate statistical discrimination: An identification problem". *Review of Economics and Statistics* 107 (3), 605–620.

BOHREN, J. A., HULL, P. and IMAS, A. (2024). *Systemic Discrimination: Theory and Measurement*. Working Paper 29820. National Bureau of Economic Research. URL: https://www.nber.org/papers/w29820.

BOHREN, J. A., IMAS, A. and ROSENBERG, M. (2019). "The Dynamics of Discrimination: Theory and Evidence". *American Economic Review* 109 (10), pp. 3395–3436.

CANAY, I. A., MOGSTAD, M. and MOUNTJOY, J. (Aug. 2023). "On the Use of Outcome Tests for Detecting Bias in Decision Making". *The Review of Economic Studies* 91 (4), pp. 2135–2167.

CHYN, E., FRANDSEN, B. and LESLIE, E. (2025). "Examiner and judge designs in economics: A practitioner's guide". *Journal of Economic Literature* 63 (2), pp. 401–439.

DEFENSORÍA DEL PUEBLO (2019). "Racismo en el Perú: Desafíos para la igualdad". Serie Informes de Adjuntía – Informe N° 008-2013-DP/ADHPD, Defensoría del Pueblo del Perú.

EMBRECHTS, P. and HOFERT, M. (2013). "A note on generalized inverses". *Mathematical Methods of Operations Research* 77, pp. 423 –432.

FEIGENBERG, B. and MILLER, C. (2022). "Would eliminating racial disparities in motor vehicle searches have efficiency costs?" *The Quarterly Journal of Economics* 137 (1), pp. 49–113.

FRANDSEN, B., LEFGREN, L. and LESLIE, E. (2023). "Judging Judge Fixed Effects". *American Economic Review* 113 (1), 253–77.

GOLDSMITH-PINKHAM, P., HULL, P. and KOLESÁR, M. (2026). "Leniency Designs: An Operator's Manual". Working paper.

GONCALVES, F. and MELLO, S. (2021). "A few bad apples? Racial bias in policing". *American Economic Review* 111 (5), pp. 1406–1441.

GURYAN, J. and CHARLES, K. K. (2013). "Taste-based or statistical discrimination: The economics of discrimination returns to its roots". *The Economic Journal* 123 (572), F417–F432.

HODERLEIN, S., HOLZMANN, H., KASY, M. and MEISTER, A. (2017). "Corrigendum: Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment". *The Review of Economic Studies* 84 (2), pp. 964–968.

HSIEH, C.-T., HURST, E., JONES, C. I. and KLENOW, P. J. (2019). "The allocation of talent and US economic growth". *Econometrica* 87 (5), pp. 1439–1474.

HULL, P. (2021). *What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making*. Working Paper 28503. National Bureau of Economic Research. URL: http://www.nber.org/papers/w28503.

HURST, E., RUBINSTEIN, Y. and SHIMIZU, K. (2024). "Task-Based Discrimination". *American Economic Review* 114 (6), pp. 1723–1768.

KLINE, P., ROSE, E. K. and WALTERS, C. R. (2022). "Systemic discrimination among large US employers". *The Quarterly Journal of Economics* 137 (4), pp. 1963–2036.

— (2024). "A discrimination report card". *American Economic Review* 114 (8), pp. 2472–2525.

LANG, K. and LEHMANN, J.-Y. K. (2012a). "Racial discrimination in the labor market: Theory and empirics". *Journal of Economic Literature* 50 (4), pp. 959–1006.

— (2012b). "Racial Discrimination in the Labor Market: Theory and Empirics". *Journal of Economic Literature* 50 (4), pp. 959–1006.

MILGROM, P. and SHANNON, C. (1994). "Monotone Comparative Statics". *Econometrica* 62 (1), pp. 157–180.

MUELLER-SMITH, M. (2015). "The criminal and labor market impacts of incarceration". Working paper.

NEUMARK, D. (2018). "Experimental research on labor market discrimination". *Journal of Economic Literature* 56 (3), pp. 799–866.

PHELPS, E. S. (1972). "The Statistical Theory of Racism and Sexism". *The American Economic Review* 62 (4), pp. 659–661.

ROSS, K. A. (2003). *Elementary Analysis: The Theory of Calculus*. 2nd. New York: Springer.

SIGSTAD, H. (2023). "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs". Working paper.

TRANSPARENCY INTERNATIONAL (2026). *Corruption Perceptions Index 2025*. Proética Peru - Transparency International.

VYTLACIL, E. (2002). "Independence, Monotonicity, and Latent Index Models: An Equivalence Result". *Econometrica* 70 (1), pp. 331–341.

Appendices

A Historical MOT with the binary esclation decision

A.1 Approach and results:

Table A.1 shows the results from implementing the historical MOT using escalation as a binary action. We estimate, separately by race, two-stage-least squares regressions of indictment ($Y = 1$) on escalation ($D = 1$) using the leniency[31] of the matched analyst as an instrument:

$$Y = \alpha_2 + \beta_2 \widehat{D} + \mathbf{X}_i \boldsymbol{\gamma}_2 + \varepsilon$$
$$D = \alpha_1 + \beta_1 \text{Leniency} + \mathbf{X}_i \boldsymbol{\gamma}_1 + \nu$$

	(1)	(2)	(3)	(4)
Panel A: European (N=9,296)				
Escalation	0.2481***	0.2321***	0.2474***	0.2302***
	(0.0182)	(0.0214)	(0.0182)	(0.0215)
Adjusted R^2	0.2677	0.2885	0.2678	0.291
Panel B: Indigenous (N=3,589)				
Escalation	0.1582***	0.1613***	0.1595***	0.1650***
	(0.0349)	(0.0349)	(0.0349)	(0.0349)
Adjusted R^2	0.1928	0.2145	0.1947	0.2274
Panel C: Difference (N=12,885)				
Escalation	0.0899**	0.0708*	0.0879*	0.0652
	(0.0391)	(0.0409)	(0.0391)	(0.0410)
Adjusted R^2	0.2539	0.2716	0.2543	0.2736
Fixed Effects	-	Region	Month	Region×Month

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A.1: RESULTS OF THE HISTORICAL MOT. POSITIVE VALUES IN PANEL C REPRESENT DISCRIMINATION AGAINST INDIGENOUS DEFENDANTS.

The columns of A.1 indicate various choices of fixed effects to represent strata within which historical analysts are assumed to be as-good-as-randomly assigned to cases. While controlling for region by month fixed effects (column (4)) represents the weakest assumption in terms of identification, these rich fixed effects increase standard errors enough to render the difference in

[31] We construct a scalar measure Z_i representing the assigned analyst's leniency (defined as propensity for escalation of report i), taking ordered values $z_0, \ldots, z_J$, where $J + 1$ is the total number of analysts in the CGR system. For instance, a value of $j = 0.5$ indicates that analyst j escalates 50% of all reports. We construct Z_i using a standard leave-out procedure capturing analysts' leniency separately by ethnicity.

Panel C insignificant at the 10% level. Given the positive magnitude of the difference for less rich sets of fixed effects, we interpret Table A.1 as providing suggestive but relatively weak evidence of taste based bias against Indigenous defendents in the historical escalation decisions at CGR.

A.2 Econometric framework for the historical MOT:

Building on Arnold et al., 2018; Hull, 2021, we can justify the 2SLS regressions above as follows.

Suppose that $\phi_{jrv}(h) = \mathbb{1}(h \geq t_{jrv})$ and $h_j(r,x) = h^*(r,x)$ for some common function h^* (i.e. beliefs are common, whether or not they are correct). Suppose that conditional on any value R, $h^*(R,X)$ is continuously distributed with a positive density such that

$$D = \mathbb{1}(h^*(R,X) \geq t_{JRV}) = \mathbb{1}(U \geq g^y_{RV}(t_{JRV}))$$

where $g^y_{RV}(t) := F_{h^*(R,X)|R,V}(t)$ and $U = g^y_{RV}(h^*(R,X))$. Notice that $U|R,V = Unif[0,1]$ by assumption and thus $U|J,R,V = Unif[0,1]$ since $J \perp\!\!\!\perp (R,X)$ by quasi-random assignment of analysts. For ease of exposition, we implicitly condition on the strata within which this assumption is maintained. The above decision rule is equivalent to the LATE monotonicity assumption holding for cases of a given race (Vytlacil, 2002).

Consequently, we can identify the quantity $g^y_{rv}(t_{jrv})$ by noticing that

$$P(D = 0|J = j, R = r, V = v) = P(U < g^y_{rv}(t_{jrv})) = g^y_{rv}(t_{jrv})$$

Define $p_{RV}(j) = P(D = 0|J = j, R, V) = g^y_{RV}(t_{jRV})$, where $p_{rv}(j)$ is the leniency of judge j among cases of race r and $V = v$. Consider now

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[Y|p_{RV}(J) = t, R, V] &= \frac{d}{dt}\mathbb{E}[Y^* \cdot D|p(J) = t, R, V] \\
&= \frac{d}{dt}\mathbb{E}[Y^* \cdot \mathbb{1}(U \geq g^y_{RV}(t_{JRV}))|g^y_{RV}(t_{JRV}) = t, R, V] \\
&= \frac{d}{dt}\mathbb{E}[Y^* \cdot \mathbb{1}(U \geq t)|R, V] \\
&= \frac{d}{dt}\int_t^1 \mathbb{E}[Y^*|U = u, R, V] \cdot du = -\mathbb{E}[Y^*|U = t, R, V] \\
&- \mathbb{E}[Y^*|U = t, R, V] = -\mathbb{E}[Y^*|h^*(R,V) = {g^y_{RV}}^{-1}(t), R, V] \\
&= -\mathbb{E}[Y^*|h^*(R,V) = {g^y_{RV}}^{-1}(t), R, V, p_{RV}(J) = t] \\
&= -\mathbb{E}[Y^*|h^*(R,V) = {g^y_{RV}}^{-1}(p_{RV}(J)), R, V, p_{RV}(J) = t] \\
&= -\mathbb{E}[Y^*|h^*(R,V) = t_{JRV}, R, V, p_{RV}(J) = t]
\end{aligned}$$

Suppose that $h^*(R, X) = \mathbb{E}[Y^*|R, X]$. Then in turn, by the law of iterated expectations:

$$\begin{aligned}
\mathbb{E}[Y^*|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t] &= \mathbb{E}\left\{\mathbb{E}[Y^*|R, X, J]|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y^*|R, X]|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t\right\} \\
&= \mathbb{E}\left\{h^*(R,X)|h^*(R,X) = t_{JRV}, R, V, p_{RV}(J) = t\right\} \\
&= \mathbb{E}\left\{t_{JRV}|h^*(R,X) = t_{JRV}, g^y_{RV}(t_{JRV}) = t, R, V\right\} \\
&= \mathbb{E}\left\{t_{JRV}|h^*(R,X) = t_{JRV} = {g^y_{RV}}^{-1}(t), R, V\right\} \\
&= \mathbb{E}[t_{JRV}|R, V]
\end{aligned}$$

Consider the case with no V for simplicity. Then the difference in MTE's:

$$\frac{d}{dt}\mathbb{E}[Y|p_i(J) = t, R = i] - \frac{d}{dt}\mathbb{E}[Y|p_e(J) = t, R = e] = \mathbb{E}[t_{Je}|R = e] - \mathbb{E}[t_{Ji}|R = i] = \mathbb{E}[t_{Je} - t_{Ji}]$$

We follow Arnold et al. (2018) in assuming that the LHS of the above expression is well-approximated by a two-stage least squares regression of Y on escalation D, instrumented by leniency $p_r(j)$ within race, and including strata fixed effects.

A.3 Leniency and monotonicity

Let D_{ij} denote whether analyst j would escalate case i. ("Unconditional") monotonicity says that for any two analysts j', j one of the analysts (j) is more "lenient" than the other in the sense that

$$D_{ij'} \geq D_{ij} \quad \text{for all cases } i$$

That is: if the more lenient analyst j would escalate the case, then so would the less-lenient analyst j' (i.e. $D_{ij} = 1 \implies D_{ij'} = 1$). Given this assumption and random assignment of analysts to cases, we can detect empirically which analyst from a pair is the more lenient, since the ordering of j' and j above is the same as the ordering of analysts based on their observed leniency $z_j = P(D_i = 0|J_i = j)$, where J_i indicates the analyst assigned to case i. To see this, use that $D_i = D_{i,J_i}$ and thus $z_j = P(D_{ij} = 0)$ and $z_j - z_{j'} = P(D_{ij} = 0 \text{ but } D_{ij'} = 1) \geq 0$, so $z_j \geq z_{j'}$.

Within-ethnicity monotonicity instead assumes that

$$D_{ij'} \geq D_{ij} \quad \text{for all cases in which } R_i = r$$

Under this assumption we instead define an analysts' leniency only using cases of a given race, i.e. $z_{jr} = P(D_i = 0|J_i = j, R_i = r)$ where R_i the ethnicity of the accused in case i. The ordering of analysts by their leniencies may differ by race. Then we say that analyst j is more lenient than analyst z' for cases of type r if $z_{jr} \geq z_{j'r}$, which follows from the above given random assignment of cases to analysts $J_i \perp\!\!\!\perp (R_i, X_i) \implies \{D_{ji} \perp\!\!\!\perp J_i\}|R_i$.

A.4 Testing the monotonicity assumption

Following Frandsen et al. (2023), we conduct a nonparametric test for the monotonicity assumption required for our instrumental variable (IV) design based on quasi-random assignment of analysts to cases. Test rejection indicates that either the exclusion restriction, the monotonicity assumption, or both, may not be upheld. Table A.2 shows the results of implementing the test. For Indigenous defendants, we fail to reject (p-value ∼1) at any confidence level, while for European defendants we fail to reject with 95% confidence but we reject with 90% confidence (p-value = 0.054).

	European	Indigenous
Test statistic	671	158
d.f.	(602)	(274)
P-value	[0.054]	[1.000]

TABLE A.2: IMPLEMENTING THE FRANDSEN ET AL. (2023) NONPARAMETRIC TEST FOR MONOTONICITY UPDATE?

A.5 Direct enumeration of monotonicity violations

Similar to Sigstad (2023), we can also test monotonicity more directly making use of the fact that in our experiment, multiple analysts see overlapping pairs of cases. We use the priority scores P_{ij} assigned to case i by analyst j. We assume that these represent an "ordinal proxy" for the counterfactual escalation decisions D_{ij} they would make, in the sense that

$$S_{ij'} > S_{ij} \implies D_{ij'} \geq D_{ij} \tag{A.1}$$

Let analysts j' and j see sets $C_j(r)$ and $C_{j'}(r)$ cases of ethnicity r, with intersection $I_{jj'}(r) = C_{jr} \cap C_{j'}(r)$. We measure violations of monotonicity separately for Indigenous and European cases (and restricting attention to cases that have not been "flipped".

For any two cases $i \neq i'$ in $I_{jj'r}$, we say that there exists a monotonicity violation if $S_{ij'} > S_{ij}$ but $S_{i'j'} < S_{i'j}$. The overall number of such violations is

$$n(r) = \sum_{\substack{j,j' \in \mathcal{J} \\ j' \neq j}} \sum_{\substack{i,i' \in I_{jj'}(r) \\ i' \neq i}} \mathbb{1}(S_{ij'} > S_{ij} \text{ and } S_{i'j'} < S_{i'j})$$

where $\mathcal{J}$ is the set of all analysts. The number of such comparisons is

$$d(r) = \sum_{\substack{j,j' \in \mathcal{J} \\ j' \neq j}} \sum_{\substack{i,i' \in I_{jj'}(r) \\ i' \neq i}} 1$$

and thus the overall rate of monotonicity violations among cases of race r can be defined as $n(r)/d(r)$. Given (A.1), this serves as an upper bound on the rate of monotoncity violations with respect to escalation D_{ij}, because the strict inequalities in S only imply weak inequalities in D. We find this rate to be 11.8% among Indigenous cases and 11.6% among European cases.

B Additional formal results

B.1 Identifying bias in accusations

The distribution of observables X over the sample of accused cases of a given race r is

$$f_{X|A=1,R=r}(x) = f_{X|R=r}(x) \cdot \frac{P(A=1|R=r,X=x)}{P(A=1|R=r)}$$

Consequently, we can identify $P(A=1|R=i,X=x) - P(A=1|R=e,X=x)/P(A=1)$ as

$$\frac{f_{X|A=1,R=i}(x)}{f_{X|R=i}(x)} \cdot \frac{P(R=i|A=1)}{P(R=i)} - \frac{f_{X|A=1,R=e}(x)}{f_{X|R=e}(x)} \cdot \frac{P(R=e|A=1)}{P(R=e)} \tag{B.1}$$

All of the quantities in (B.1) are in principle identified, given data on X throughout the general population of potential cases (e.g. government workers). Even if we did not know the unconditional probability of accusation $P(A=1)$, it is positive and hence we can infer the sign of $P(A=1|R=i,X=x) - P(A=1|R=e,X=x)$, the difference in accusation-rates among observably-equivalent Indigenous and European cases.

B.2 A pooled test for ERM violations averaged across analysts

Note that by the law of iterated expectations over analysts J and over case observables V, Corollary 1 implies that $\mathbb{E}[H|R=e,F=1] = \mathbb{E}[g^h_{JV}(i,S)|R=e,F=1]$ and $\mathbb{E}[H|R=i,F=1] = \mathbb{E}[g^h_{JV}(e,S)|R=i,F=1]$. Thus in implementing an "averaged" test of ERM, we only need to form regression predictions $g^h_{JV}(r,S) = \mathbb{E}[H|R=r,J,S,V,F=0]$ for each r and test that $H - g^h_{JV}(r,S)$ is mean zero among the sample of flipped cases for the opposite ethnicity. For ease of implementation, we assume that $g^h_{JV}(R,S) := \mathbb{E}[H|R,J,S,V,F=0]$ is linear in S and its square, V, and analyst-specific fixed effects, with all coefficients allowed to vary by R. This leads to a fixed-effects regression of the form $H = \alpha_{JR} + \beta_{0R}S + \beta_{1R}S^2 + \gamma'_R V + \epsilon$ where $\mathbb{E}[\epsilon|J,R,S,V,F=0] = 0$. This equation is estimated separately by race, and predicted values formed across all observations. Results are presented in Table B.1.

B.3 A systematic search over v satisfying ERM

Let $\mathcal{X}$ denote the set of non-race variables visible to the analyst, and let $\mathcal{V}^* \subseteq \mathcal{X}$ be the "true" subset of these variables that enter into analysts decisions of S.

	(1)	(2)	(3)	(4)
Moment value (European cases)	0.0278***	0.0284***	0.0260***	0.0277***
	(0.00384)	(0.00394)	(0.00679)	(0.00702)
Moment value (Indigenous cases)	-0.0184***	-0.0167***	-0.0172*	-0.0175*
	(0.00418)	(0.00364)	(0.00676)	(0.00818)
Typology controls		X		X
Department controls			X	X
Number of obs.	19089	19089	19089	19089

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE B.1: TESTS OF ASSUMPTION ERM VIA THE OVERIDENTIFICATION RESTRICTION IN PROPOSITION 1. COLUMNS CORRESPOND TO VARIOUS CHOICES OF V: EITHER INDICATORS FOR CASE TYPOLOGY OR FOR GEOGRAPHICAL INDICATORS (PERUVIAN *departamentos*). UNDER ERM FOR A GIVEN CHOICE OF V, THE VALUE OF THE MOMENT SHOULD BE ZERO FOR EACH CASE RACE.

We've seen above that ERM would hold trivially if $\mathcal{V} = \mathcal{X}$, the full set of case observables. Note that if ERM holds with $\mathcal{V}$, then it also holds for any larger set of variables $\mathcal{V} \subset \mathcal{V}'$, because the function $\phi_{jrv'}$ need not depend on the components of v' that are not in v. Thus, the assumption that ERM holds with $\mathcal{V} = \emptyset$ (analagous to the assumption of Canay et al. (2023) but for a continuous action) is the strongest version of ERM possible. For concreteness, we assume that Assumption ERM holds for no strict subsets of $\mathcal{V}^*$, i.e. it is a *minimal* subset of $\mathcal{X}$ that makes Assumption ERM hold.

Suppose that for any $\mathcal{V} \subset \mathcal{X}$: $g^h_{jv}(r,s) = \alpha^{\mathcal{V}}_{jr} + \beta^{\mathcal{V}}_{0r} s + \beta^{\mathcal{V}}_{1r} s^2 + \gamma^{\mathcal{V}\prime}_r v$, where note that all coefficients can depend on the set $\mathcal{V}$. Let $g^*_{jv}(r,s)$ be the function $g^h_{jv}(r,s)$ when $\mathcal{V} = \mathcal{V}^*$, and denote its coefficients as $g^*_{jv}(r,s) = \alpha^*_{jr} + \beta^*_{0r} s + \beta^*_{1r} s^2 + \gamma^{*\prime}_r v$

Lemma 3. *$\lambda_j = 0$ for any $v_j \notin \mathcal{V}^*$.*

Proof. Let v^* be the part of v that corresponds to $\mathcal{V}^*$

$$
\begin{aligned}
g^h_{jv}(r,s) &= \mathbb{E}[H|R=r, J=j, S=s, V=v, F=0] \\
&= \mathbb{E}[h_j(r,X) + U|R=r, J=j, S=s, V=v, F=0] \\
&= \mathbb{E}[h_j(r,X)|R=r, J=j, s_j(r,X)=s, V=v, F=0] \\
&= \mathbb{E}[h_j(r,X)|R=r, J=j, s_j(r,X)=s, V^*=v^*, V^c=v^c] \\
&= \mathbb{E}[h_j(r,X)|R=r, J=j, h_j(r,X)=\phi^{-1}_{jrv^*}(s), V^*=v^*, V^c=v^c] = \phi^{-1}_{jrv^*}
\end{aligned}
$$

which does not vary with v^c. □

C Experimental Design

During July and August 2021, 182 analysts from CGR (approximately 90% of the triaging staff) evaluated 196 whistleblower reports previously assessed by CGR. Analysts assigned scores from 0 to 100, indicating the priority and likelihood of indictment, with 100 representing the highest priority. The experiment was conducted virtually following CGR Covid-19 safety protocols. Figure C.1 shows the experiment's interface.

Sample Selection. Reports were selected based on specific criteria: (i) text length between 50 and 350 words; (ii) issued between 2018 and 2020; (iii) allegations involving illicit enrichment, favoritism, or deficient public works execution; (iv) defendants identified as municipal mayors; and (v) defendant's full name explicitly mentioned. The final sample aimed to reflect Peru's ethnic composition.

Participants. Analysts were predominantly male, aged 30-40, university-educated, primarily trained in Law or Engineering, and employed at various regional CGR offices. Participation was mandated by CGR management, with analysts informed their evaluations would assist in designing a prioritization algorithm in collaboration with the Inter-American Development Bank and Columbia University. Analysts were unaware that surnames had been altered in some reports.

Randomization Procedure. To experimentally manipulate the apparent ethnicity of a case, we randomly switched defendant surnames between Indigenous (Quechua and Aymara) and European (Spanish) origins using regional surname dictionaries derived from historical CGR data and Artiles (2024). The randomization was stratified by defendant ethnicity (Indigenous or European) and case outcome (dismissed at escalation, escalated but dismissed at indictment, or indicted). Analysts were then assigned a random sequence of cases with the same stratification and ensuring that no analysts saw cases from their own regional office to prevent familiarity with the cases.[32]

[32]The detailed randomization algorithm is presented in appendix G.

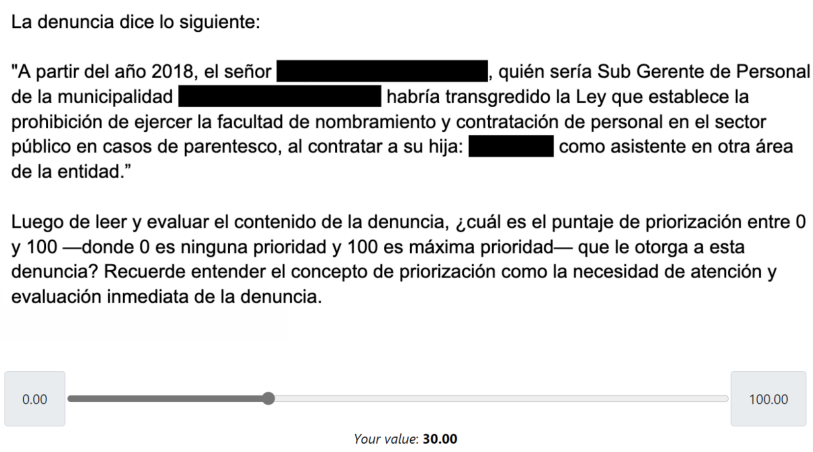


FIGURE C.1: EXPERIMENT INTERFACE

D Indictment Prediction Algorithm

A key component of the decomposition derived in section **??**, with results presented in section 5.2 is the prediction $\hat{h}(R, X)$ of the probability that a case with observables X and race R leads to an indictment. To make these predictions, we develop the prediction algorithm described in this appendix.

We proceed in three steps:

1. Take the full historical database, spanning 2013 until the end of 2021 and build a set of 1,087 potential features that could be used for the prediction task.
2. Bootstrap to reduce the 1,087 features down to the 59 most predictive variables.
3. Use these 59 variables in a range of models and validate them.

We discuss each of these steps in turn below.

D.1 Building Potential Features

Our dataset contains the universe of 69,751 whistleblower complaints from 2013 until the end of 2021.[33] We combed through the database and identified 312 columns that we might potentially be able to use for prediction. From this initial set we we take 101 features to use for the predictions. The features we use are summarized in table XXX

Starting from the 101 variables, we perform a range of transformations in a column-wise fashion to allow for rich patterns of nonlinearity and interactions in our predictive model. The transformations consist in functions that return a same-length vector per input column. They are organized into two families: (i) *stage-agnostic* functions, and (ii) *stage-dependent* functions, which perform aggregating column behaviour across process stages per case. In total there are **83** distinct transforms: stage-agnostic **33** (categorical 9, numeric 9, text 11, date 4) and stage-dependent **50** (categorical 8, numeric 20, text 11, date 11). A simple text detector routes `Utf8` columns either to text or categorical pipelines.

Stage-agnostic functions. These are dispatched by `original_functions_to_dataset` based on dtype (`Utf8`, `Date`, `Float64/Int32/Int64`).

- `cat_one_hot` — One-hot encode top levels (others bucket). *Applies to: categorical*.
- `cat_woe` — Weight-of-Evidence (per level). *Categorical*.
- `cat_tr` — Target mean encoding (per level). *Categorical*.
- `cat_ordering_by_tr` — Rank of levels by target rate. *Categorical*.

[33] In January 2022 CGR migrated to a new IT platform with a slightly different database structure.

- `cat_frequency` — Relative frequency of the level. *Categorical*.
- `cat_ordinal` — Ordinal code by frequency order. *Categorical*.
- `cat_log_odss` — Univariate logistic log-odds from dummies. *Categorical*.
- `cat_iv` — Information Value (IV) per level. *Categorical*.
- `cat_label` — Count of occurrences of the level. *Categorical*.
- `num_quantile_binning_by_median` — Quantile bins; median per bin. *Numeric*.
- `num_quantile_binning_by_tr` — Quantile bins; target rate per bin. *Numeric*.
- `num_k_means_binning` — K-means discretization (best silhouette). *Numeric*.
- `num_uniform_binning_by_median` — Equal-width bins; median per bin. *Numeric*.
- `num_uniform_binning_by_tr` — Equal-width bins; target rate per bin. *Numeric*.
- `num_woe_binning` — WoE optimal binning/cuts. *Numeric*.
- `num_log_odss` — Univariate logistic log-odds. *Numeric*.
- `num_normalization` — Z-score normalization. *Numeric*.
- `num_median_normalization` — Robust scaling by median/IQR. *Numeric*.
- `txt_conteo_words` — Word count. *Text*.
- `txt_conteo_characters` — Character count. *Text*.
- `txt_conteo_sentences` — Sentence count. *Text*.
- `txt_conteo_syllable` — Syllable count. *Text*.
- `txt_avg_words_by_sentence` — Average words per sentence. *Text*.
- `txt_avg_syllables_by_sentence` — Average syllables per sentence. *Text*.
- `txt_tf_idf_ten` — TF-IDF (top-10 terms). *Text*.
- `txt_coleman_liau_readability` — Coleman–Liau index. *Text*.
- `txt_flesch_kincaid_readability` — Flesch–Kincaid score. *Text*.
- `txt_beto_embedding` — BETO (Spanish BERT) embedding (mean-pooled). *Text*.
- `txt_label` — Frequency rank code for tokens. *Text*.
- `date_year` — Year extracted from date. *Date*.

- `date_month` — Month number extracted from date. *Date*.
- `date_date` — Date (drop time component). *Date*.
- `date_ratio_last_time_span_fecha_inicio_fin` — Span between first and last observed date for the id. *Date*.

Stage-dependent functions. These are orchestrated by `transaccional_functions_to_dataset`. They compute per-case summaries across stages (e.g., dispersion, ratios, WoE/IV/log-odds aggregates, and temporal intervals).

- `trx_cat_conteo` — # distinct levels across stages (per id). *Categorical*.
- `trx_cat_entropia` — Entropy of level distribution across stages. *Categorical*.
- `trx_cat_frequency` — Relative frequency across stages. *Categorical*.
- `trx_cat_woe` — Stage-level WoE summary (per id). *Categorical*.
- `trx_cat_log_odss` — Stage-level log-odds summary. *Categorical*.
- `trx_cat_iv` — Stage-level IV summary. *Categorical*.
- `trx_cat_tr` — Stage-level target rate summary. *Categorical*.
- `trx_cat_ordering_by_tr` — Stage-level rank by target rate. *Categorical*.
- `trx_num_uniform_binning_by_median` — Stage-agg of equal-width bin medians. *Numeric*.
- `trx_num_uniform_binning_by_tr` — Stage-agg of equal-width bin target rates. *Numeric*.
- `trx_num_quantile_binning_by_median` — Stage-agg of quantile bin medians. *Numeric*.
- `trx_num_quantile_binning_by_tr` — Stage-agg of quantile bin target rates. *Numeric*.
- `trx_num_k_means_binning` — Stage-agg of k-means discretization. *Numeric*.
- `trx_num_woe_binning` — Stage-agg of WoE bins. *Numeric*.
- `trx_log_odss` — Stage-agg of univariate logistic log-odds. *Numeric*.
- `trx_num_normalization_with_avg` — Stage-agg of z-score normalization. *Numeric*.
- `trx_num_normalizacion_with_std` — Stage-agg of std-score normalization. *Numeric*.
- `trx_num_normalization_with_median` — Stage-agg of robust (median/IQR) scaling. *Numeric*.
- `trx_num_promedio` — Mean across stages (per id). *Numeric*.

- `trx_num_varianza` — Variance across stages. *Numeric*.
- `trx_num_ratio_last_obs_vs_avg` — Last value / stage mean. *Numeric*.
- `trx_num_desviacion_estandar` — Standard deviation across stages. *Numeric*.
- `trx_num_mediana` — Median across stages. *Numeric*.
- `trx_num_minimo` — Minimum across stages. *Numeric*.
- `trx_num_max` — Maximum across stages. *Numeric*.
- `trx_num_ratio_max_min` — Max/Min ratio across stages. *Numeric*.
- `trx_num_ratio_max_mediana` — Max/Median ratio across stages. *Numeric*.
- `trx_num_ratio_max_promedio` — Max/Mean ratio across stages. *Numeric*.
- `trx_txt_target` — Stage-level target mean encoding. *Text*.
- `trx_txt_mean` — Stage-level mean encoding (duplicate defs; treat as one). *Text*.
- `trx_txt_conteo_characters_varianza` — Variance of character count across stages. *Text*.
- `trx_txt_conteo_words_varianza` — Variance of word count across stages. *Text*.
- `trx_txt_conteo_sentences_varianza` — Variance of sentence count across stages. *Text*.
- `trx_txt_conteo_syllable_varianza` — Variance of syllable count across stages. *Text*.
- `trx_txt_avg_words_by_sentence_with_var` — Variance of avg words/sentence across stages. *Text*.
- `trx_txt_avg_syllables_by_setence_var` — Variance of avg syllables/sentence across stages. *Text*.
- `trx_txt_tf_idf_ten_var` — Variance of TF-IDF top-10 representation across stages. *Text*.
- `trx_txt_coleman_liau_readability_var` — Variance of Coleman–Liau index. *Text*.
- `trx_txt_flesch_kincaid_readability_var` — Variance of Flesch–Kincaid. *Text*.
- `trx_date_year` — Year of last observed date (per id). *Date*.
- `trx_date_month` — Month of last observed date. *Date*.
- `trx_date_date` — Last observed date (per id). *Date*.
- `trx_date_time_span` — Total time span (sum of diffs) across stages. *Date*.

- `trx_date_ratio_last_time_span_average_time_span` — Last interval / average interval. *Date*.
- `trx_date_average_time_span` — Average interval between consecutive dates. *Date*.
- `trx_date_min_time_span` — Minimum interval between dates. *Date*.
- `trx_date_max_time_span` — Maximum interval between dates. *Date*.
- `trx_date_std_time_span` — Standard deviation of intervals. *Date*.
- `trx_date_ratio_max_time_span_min_time_span` — Max/Min interval ratio. *Date*.
- `trx_date_antiguedad` — Age since last observed date (days). *Date*.

After these transformations, we arrive at a dataset of 1,087 features to be potentially used for the predictions.

D.2 Bootstrapping to Select Features

Given the size of the dataset and the number of potential features, computational constraints make it desirable to whittle down the set of features to a more manageable number.

We do this in four steps.

First, we take the 1,087 features we have created and group them into XXXX clusters based on how correlated they are with each other. To do this, we compute the correlation between each pair of variables and use this as the distance measure in a K-nearest-neighbors clustering algorithm. Figure XXXX shows the algorithm is highly successful. The figure shows a heatmap of the correlations between the variables where the variables are sorted by the cluster they belong to, and we can see that the heatmap displays strong blocking along the diagonal.

Second, we run 100,000 iterations of the following algorithm:

1. randomly select 50 of the XXXX clusters
2. randomly select one variable from each of the 50 clusters.
3. train a random forest algorithm using these 50 variables
4. compute the gini score of each of the 50 variables.

Third, for each variable, we compute the average of their gini score across all of the bootstrap iterations in which the variable was selected. We then select the XXX variables from each cluster with the highest average gini scores. This resulted in a list of 74 variables to include in the model.

Fourth, we used all 74 variables in a LASSO model and use the 59 variables that the LASSO with the MSE0minimizing + 1 standard deviation penalty returned. The 59 variables are listed in table XXXX.

D.3 Training and Validating Predictions

With our final set of 59 variables, we trained six classification models.

1. LASSO model with the 9 MSE-minimizing variables.
2. Logit model with the 59 variables.
3. Random forest model with the 59 variables.
4. Random forest model using the top 20 variables.
5. LASSO model with the top 20 variables.
6. XGBoost model using the 59 variables.

Each model was trained on 70% of the data and then tested on the remaining 30%. We trace out the Receiver Operating Curve (ROC) for each model and compute the Area Under the Curve (AUC) and GINI coefficient in both the training and the test data. Figure D.1 shows the results. All six models perform well, but the {**MB:** *WHICH?*} model is our preferred model because {**MB:** *WHY?*}.

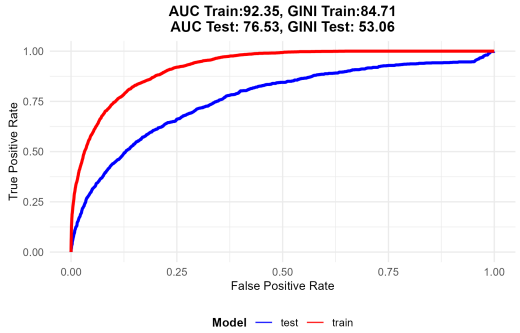
Based on this analysis we selected the XXX model for our prediction exercise. We retrained the model using the full dataset and adding the ethnicity of the defendant as an additional feature.[34] From this model we form three predictions for each case

1. The predicted probability of indictment of the case with a European name.
2. The predicted probability of indictment of the case with an Indigenous name.
3. The predicted probability of indictment of the case without a name.

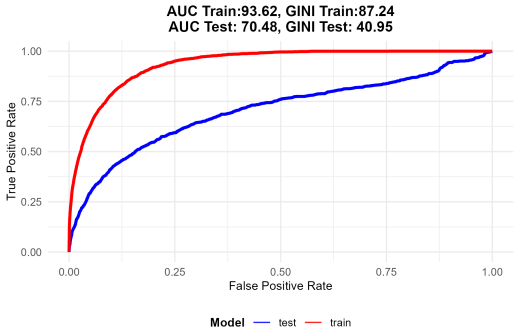
[34] See Appendix E for details on how we extracted the ethnicity of the defendants from the text of the complaints.

FIGURE D.1: VALIDATING CLASSIFICATION MODELS

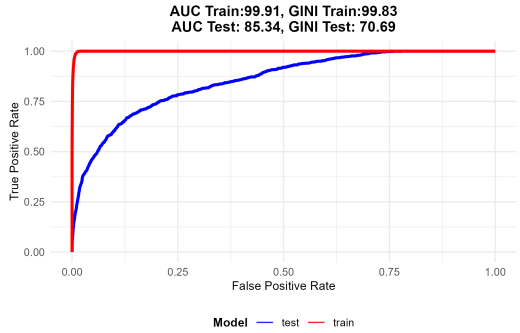
PANEL A: 9-VAR LASSO



PANEL B: 59-VAR LASSO



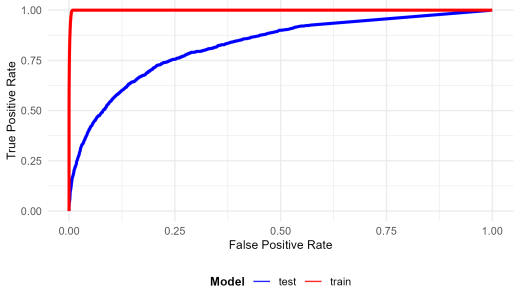
PANEL C: 59-VAR RANDOM FOREST



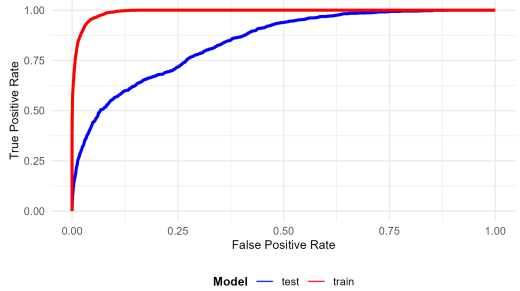
PANEL D: XGBOOST BASELINE 9-VARS



PANEL E: 9-VAR RANDOM FOREST



PANEL F: XGBOOST BASELINE 59-VARS



Notes: There are notes here.

E Assigning Ethnicities to Defendants

This appendix describes how we extract names and assign them ethnicities.

1. We start with the LLM extracting names. Here's how it works and the benchmarking that Diego did.
2. We start from Miriam's algo.
3. For the ones that Miriam's algo doesn't catch (1,000 names), we use Ale's procedure. need to check what that was. and LLM-based thing but what?
4. we randomly check XXXX of these and fix any mistakes, but the mistake rate was low YYY%.

F Proofs

F.1 Proof of Proposition 1

Since ϕ_{jev} is increasing and right-continuous, it follows that $\phi_{jev}(h_j(e,x)+\delta) \leq \phi_{jiv}(h_j(i,x))$ if and only if $h_j(e,x)+\delta \leq \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x)))$ (see e.g. item 5 of Proposition 2.3 in Embrechts and Hofert 2013 for a proof). Thus

$$\delta_{jex} = \sup\{\delta : h_j(e,x)+\delta \leq \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x)))\} = \phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(e,x)$$

The case of δ_{jix} is symmetric.

F.2 Proof of Lemma 1

$$\begin{aligned} g^h_{jv}(r,s) &= \mathbb{E}[h_J(\tilde{R},X)+U | R=r, \phi_{J\tilde{R}V}(h_J(\tilde{R},X))=s, J=j, V=v, F=0] \\ &= \mathbb{E}[h_j(r,X) | R=r, h_j(r,X)=\phi_{jrv}^{-1}(s), J=j, V=v, F=0] \\ &\qquad + \mathbb{E}[U | R=r, h_j(r,X)=\phi_{jrv}^{-1}(s), J=j, V=v, F=0] \\ &= \phi_{jrv}^{-1}(s) + \mathbb{E}[U | R=r, J=j, F=0] = \phi_{jrv}^{-1}(s) + 0 \end{aligned}$$

where we have used ERM in the first equality, and CLASSICAL in the final line. Note that CLASSICAL also implies that $g^h_{jbv}(r,s) = \phi_{jrv}^{-1}(s)$ if we define $g^h_{jvb}(r,s) := \mathbb{E}[H | R=r, S=s, J=j, B=b, V=v, F=0]$, implying that $g^h_{jbv}(r,s)$ does not depend on b given ERM and CLASSICAL. We use this latter definition of g^h in estimation, as a fixed effects regression with fixed effects for the pair J, B.

Technical note: Given that the event $\phi_{J\tilde{R}V}(h_J(\tilde{R},X)) = s$ has probability zero, some care is required in replacing the conditional expectation $\mathbb{E}[h_j(r,X) | R=r, \phi_{jrv}(h_j(r,X))=s, J=j, V=$

$v, F = 0]$ with $\mathbb{E}[h_j(r, X)|R = r, h_j(r, X) = \phi_{jrv}^{-1}(s), J = j, V = v, F = 0]$. The reason is that although $\phi_{jrv}(h_j(r, X)) = s$ occurs if an only if $h_j(r, X) = \phi_{jrv}^{-1}(s)$, conditional distributions that condition on $\phi_{jrv}(h_j(r, X)) = s$ are defined from a shrinking sequence of intervals $[s, s + \epsilon]$ as $\epsilon \downarrow 0$, while conditional distributions that condition on $h_j(r, X) = \phi_{jrv}^{-1}(s)$ are defined from a shrinking sequence of intervals $[\phi_{jrv}^{-1}(s), \phi_{jrv}^{-1}(s) + \epsilon]$ as $\epsilon \downarrow 0$. If instead of holding $V = v$ fixed, we held $X = x$ fixed, the two conditional expectations would not need to be the same for multivariate V (HODERLEIN et al., 2017).

F.3 Proof of Theorem 1

Consider item 1. Since $\{F \perp\!\!\!\perp X\}|J, B$ by RAND, we have that $S|R = e, J, B, F = 0 \sim s_J(e, X)|J, B$ and $S|R = e, J, B, F = 1 \sim s_J(i, X)|J, B$, cf. the definition of counterfactual outcomes in Section **??**. Then by ERM we have that $s_j(r, x) = \phi_{jvr}(h_j(r, x))$ and thus

$$\begin{aligned}
&\mathbb{E}[g_{JV}^h(i, S)|R = e, J, B, F = 1] - \mathbb{E}[g_{JV}^h(e, S)|R = e, J, B, F = 0] \\
&= \mathbb{E}[\phi_{JiV}^{-1}(\phi_{JiV}(h(i, X)))|R = e, J, B] - \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h(e, X)))|R = e, J, B] \\
&= \mathbb{E}[h(i, X) - h(e, X)|R = e, J, B] \qquad \text{(F.1)}
\end{aligned}$$

using Lemma 1 to substitute ϕ_{jvr}^{-1} for $g_{jv}^h(r, s)$. Then

$$BBB_e = \mathbb{E}\left\{\mathbb{E}[h(i, X) - h(e, X)|R = e, J, B] \mid R = e\right\}$$

by the law of iterated expectations. The proof for BBB_i (item 3) is analagous.

For item 2, we have similarly that:

$$\begin{aligned}
&\mathbb{E}[g_{JV}^h(e, S)|R = e, J, B, F = 1] - \mathbb{E}[g_{JV}^h(e, S)|R = e, J, B, F = 0] \\
&= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h(i, X)))|R = e, J, B] - \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h(e, X)))|R = e, J, B] \\
&= \mathbb{E}[\phi_{JeV}^{-1}(\phi_{JiV}(h(i, X))) - h(e, X)|R = e, J, B] \\
&= \mathbb{E}[\underbrace{\phi_{JeV}^{-1}(\phi_{JiV}(h(i, X))) - h(i, X)}_{\text{taste-based bias}}|R = e, J, B] + \mathbb{E}[\underbrace{h(i, X) - h(e, X)}_{\text{belief-based bias}}|R = e, J, B]
\end{aligned}$$

which yields TOT_e by the law of iterated expectations over $J, B|R = e$. The proof for TOT_i (item 4) is analagous.

F.4 Proof of Lemma 2

Given that S is pinned down conditional on X, R, J and F:

$$\begin{aligned}
p_{jv}(r,s) &= \mathbb{E}[Y|R=r,S=s,J=j,V=v,F=0] \\
&= \mathbb{E}\{\mathbb{E}[Y|X,R=r,S=s,J=j,V=v]|R=r,S=s,J=j,V=v,F=0\} \\
&= \mathbb{E}\{\mathbb{E}[Y|X,R=r,\cancel{\phi_{jrv}(h_j(r,X))=s},J=j,F=0]|R=r,S=s,J=j,V=v,F=0\} \\
&= \mathbb{E}\{\mathbb{E}[Y|X,R=r,J=j,F=0]|R=r,S=s,J=j,V=v,F=0\}
\end{aligned}$$

The random variable within the outer expectation is

$$\begin{aligned}
&\mathbb{E}[Y|X,R=r,J=j,F=0] = \mathbb{E}[Y|X,R=r,J=j,F=0] = h(r,x) \\
&= h_j(r,x) - \{h_j(r,x) - h(r,x)\}
\end{aligned}$$

This establishes that

$$\begin{aligned}
p_{jv}(r,s) &= g^*_{jv}(r,s) - \mathbb{E}[h_j(r,X) - h(r,X)|\tilde{R}=r,S=s,J=j,V=v,F=0] \\
&= g^*_{jv}(r,s) - \mathbb{E}[h_j(r,X) - h(r,X)|R=r,\phi_{jrv}(h_j(r,X))=s,J=j,V=v,F=0]
\end{aligned}$$

where we let $g^*_{jv}(r,s) := \mathbb{E}[h_J(R,X)|R=r,S=s,J=j,V=v,F=0]$. This represents a special case of the quantity $g^h_{jv}(r,s)$, when $U=0$ (so that $H := h_J(\tilde{R},X)$) with probability one. In this case Assumption CLASSICAL would apply trivially, and thus we can use the steps of Lemma 1 to establish that $g^*_{jv}(r,s) = \phi^{-1}_{jrv}(s)$.

F.5 Proof of Theorem 2

For the equation targeting TOT_e, we have using Lemma 2 that:

$$\begin{aligned}
&\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=1] - \mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=0] \\
&= \mathbb{E}[\phi^{-1}_{JeV}(\phi_{JiV}(h_J(i,X)))|R=e,J,B] - \mathbb{E}[\phi^{-1}_{JeV}(\phi_{JeV}(h_J(e,X)))|R=e,J,B] + W^{TOT}_e(J,B) \\
&= \mathbb{E}[\phi^{-1}_{JeV}(\phi_{JiV}(h_J(i,X))) - h_J(e,X)|R=e,J,B] + W^{TOT}_e(J,B) \\
&= \mathbb{E}[\underbrace{h_J(i,X) - h_J(e,X)}_{\text{belief-based bias}}|R=e,J,B] + \mathbb{E}[\underbrace{\phi^{-1}_{JeV}(\phi_{JiV}(h_J(i,X))) - h_J(i,X)}_{\text{taste-based bias}}|R=e,J,B] + W^{TOT}_e(J,B)
\end{aligned}$$

by steps analogous to those in Theorem 1, where the term $W_e^{TOT}(J,B)$ is

$$\begin{aligned} W_e^{TOT}(J,B) &:= \mathbb{E}\{\mathbb{E}[\tilde{h}_J(e,X)|R=e,S,J,V,F=0]|R=e,J,B,F=0\} \\ &\qquad - \mathbb{E}\{\mathbb{E}[\tilde{h}_j(e,X)|R=e,S,J,V,F=0]|R=e,J,B,F=1\} \\ &= \int dF_{X|R=e,J,B}(x)\cdot\{\mathbb{E}[\tilde{h}_J(e,X)|R=e,S=s_J(e,x),J,V=v,F=0] \\ &\qquad - \mathbb{E}[\tilde{h}_J(e,X)|R=e,S=s_J(i,x),J,V=v,F=0]\} \end{aligned} \tag{F.2}$$

using RAND.

Note that this term differs from zero only insofar as the distribution of $s_j(e,X)=\phi_{jeV}(h_j(e,X))$ differs from that of $s_j(e,X)=\phi_{jiV}(h_j(i,X))$, with each being correlated with $\tilde{h}_j(e,X)$ (conditional on V and not being flipped). We might expect this quantity to be small in settings with simple patterns of heterogeneity. The final result then follows with $W_r^{TOT}=\mathbb{E}[W_r^{TOT}(J,B)|R=r]$, and is analagous for TOT_i.

Similarly, for the BBB_e expression, we obtain:

$$\begin{aligned} &\mathbb{E}[p_{JV}(i,S)|R=e,J,B,F=1]-\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=0] \\ &=\mathbb{E}[\phi_{JiV}^{-1}(\phi_{JiV}(h_J(i,X)))|R=e,J,B,]-\mathbb{E}[\phi_{JeV}^{-1}(\phi_{JeV}(h_J(e,X)))|R=e,J,B,]+W_e^{\mathcal{B}}(J,B) \\ &=\mathbb{E}[\underbrace{h_J(i,X)-h_J(e,X)}_{\text{belief-based bias}}|R=e,J,B,]+W_e^{\mathcal{B}}(J,B) \end{aligned}$$

where the term $W_e^{\mathcal{B}}(J,B)$ is

$$\begin{aligned} W_e^{\mathcal{B}}(J,B) &= \mathbb{E}\{\mathbb{E}[\tilde{h}_J(e,X)|R=e,S,J=j,V,F=0]|R=e,J,B,F=0\} \\ &\qquad - \mathbb{E}\{\mathbb{E}[\tilde{h}_J(i,X)|R=i,S,J=j,V,F=0]|R=e,J,F=1\} \\ &= \int dF_{X|R=e,J,B}(x)\cdot\{\mathbb{E}[\tilde{h}_J(e,X)|R=e,S=s_J(e,x),J,V=v,F=0] \\ &\qquad - \mathbb{E}[\tilde{h}_J(i,X)|R=i,S=s_J(i,x),J,V=v,F=0]\} \end{aligned}$$

again using RAND. Define $W_r^{\mathcal{B}}=\mathbb{E}[W_r^{\mathcal{B}}(J,B)|R=r]$, and the argument is analagous for BBB_i. Note that $W_r^{\mathcal{B}}$ need not have the opposite sign as IBB_r. Despite the latter having means of $\tilde{h}_J(i,X)$ showing up with a minus sign and means of $\tilde{h}_J(e,X)$ showing up with a plus sign (the opposite of $W_r^{\mathcal{B}}$), the expression above for $W_r^{\mathcal{B}}$ involves conditioning on S while the definition of IBB_r does not.

If to estimate TBB_e we use the difference:

$$\begin{aligned} &\mathbb{E}\left\{\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=1]|R=e\right\}-\mathbb{E}\left\{\mathbb{E}[p_{JV}(i,S)|R=e,J,B,F=1]|R=e\right\} \\ &=\mathbb{E}\left\{\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=1]-\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=0]|R=e\right\} \\ &\qquad -\mathbb{E}\left\{\mathbb{E}[p_{JV}(i,S)|R=e,J,B,F=1]-\mathbb{E}[p_{JV}(e,S)|R=e,J,B,F=0]|R=e\right\}, \end{aligned}$$

we obtain by the above calculations $TOT_e - BBB_e + (W_e^{TOT} - W_e^{\mathcal{B}})$, where the researcher error W_e^{TB} from the desired quantity TBB_e is the average of

$$\int dF_{X|R=e,J,B}(x) \cdot \{\mathbb{E}[\tilde{h}_J(i,X)|R=i, S=s_j(i,x), J, V=v, F=0] \\ - \mathbb{E}[\tilde{h}_J(e,X)|R=e, S=s_j(i,x), J, V=v, F=0]\}$$

over the distribution of J, B conditional on $R=e$. This can be expected to be close to $-W_e^{\mathcal{B}}$, if W_e^{TOT} is small (as we might expect, by the given above).

F.6 Proof of Proposition 2

Let $g^h_{jvw}(r,S) := \mathbb{E}[H|\tilde{R}=r, S, J=j, V=v, W=w]$. Then

$$\begin{aligned}
&\mathbb{E}[g^h_{jvw'}(r,S)|\tilde{R}=r, J=j, V=v, W=w] = \mathbb{E}[g^y_{jvw'}(r, s_j(r,X))|\tilde{R}=r, J=j, V=v, W=w] \\
&= \int dF_{X|rjvw}(x) \cdot g^h_{jvw'}(r, s_j(r,x)) \\
&= \int dF_{X|rjvw}(x) \cdot \mathbb{E}[h_j(r,X)+U|\tilde{R}=r, J=j, s_j(r,X)=s_j(r,x), V=v, W=w'] \\
&= \int dF_{X|rjw}(x) \cdot h_j(r,x) + \int dF_{X|rjvw}(x) \cdot \mathbb{E}[U|\tilde{R}=r, J=j, s_j(r,X)=s_j(r,x), V=v, W=w'] \\
&= \mathbb{E}[h_j(r,X)|R=e, J=j, V=v, W=w] + 0 \\
&= \mathbb{E}[H|\tilde{R}=r, J=j, V=v, W=w]
\end{aligned}$$

Direct proof in the lab experiment setting

Letting $g^h_{jv}(r,s) := \mathbb{E}[H|R=r, S=s, J=j, V=v, F=0]$:

$$\begin{aligned}
&\mathbb{E}[g^h_{jv}(i,S)|R=e, J=j, V=v, F=1] = \mathbb{E}[g^h_{jv}(i, s_j(i,X))|R=e, J=j, V=v, F=1] \\
&= \int dF_{X|ejv1}(x) \cdot g^h_{jv}(i, s_j(i,x)) \\
&= \int dF_{X|ejv1}(x) \cdot \mathbb{E}[h_j(i,X)+U|R=i, J=j, s_j(i,X)=s_j(i,x), V=v, F=0] \\
&= \int dF_{X|ej1}(x) \cdot h_j(i,x) + \int dF_{X|ejv1}(x) \cdot \mathbb{E}[U|R=i, J=j, s_j(i,X)=s_j(i,x), V=v, F=0] \\
&= \mathbb{E}[h_j(i,X)|R=e, J=j, V=v, F=1] + 0 = \mathbb{E}[H|R=e, J=j, V=v, F=1]
\end{aligned}$$

where we use (13) in the fourth step to replace $\mathbb{E}[h_j(i,X)|R=i, J=j, s_j(i,X)=s_j(i,x), V=v, F=0]$ with $h_j(i,x)$, because (13) and that X and x share a v means that with probability one $s_j(i,X)=s_j(i,x)$ implies that $h_j(i,X)=h_j(i,x)$. The proof for Indigenous cases is identical.

G Randomization algorithm

Our randomization procedure proceeds as detailed explicitly below. The main steps are:

1. Each denuncia is assigned a *stratum*, based on it's outcome (not escalated, escalated but not indicted, indicted), and the race of the defendent
2. For each analyst, we form *blocks* of cases that analysts can be assigned to see. Cases in their own regional office are excluded from their blocks. The probability that a block will be assigned to analyst j's block depends on the stratum to which it was assigned.
3. A random order of presentation for the cases in analyst j's block is determined
4. Name switches determined within block, with probability depending on defendant race

Algorithm 1: Randomization algorithm

Input: A set D of 196 complaints with attributes `type` and `ind`,
A set A of 204 analysts, each with an exclusion flag $\texttt{exclude}_a(d) \in \{0, 1\}$.

Output: For each analyst $a \in A$, for each $d \in D$:

- $\texttt{showing_order}_a(d)$ (the random display order)
- $\texttt{artificial_denuncia}_a(d) \in \{0, 1\}$ (whether the name is switched)

```
// 1. Stratification of all complaints
```

foreach $d \in D$ **do**

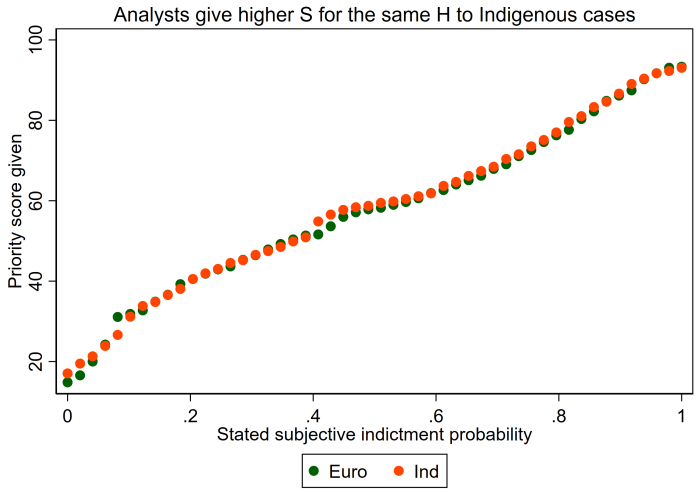
  if *type[d]=1 and ind[d]=0* **then**
    stratum[d]$\leftarrow 1$ // not escalated-euro
  else if *type[d]=1 and ind[d]=1* **then**
    stratum[d]$\leftarrow 2$ // not escalated-ind
  else if *type[d]=2 and ind[d]=0* **then**
    stratum[d]$\leftarrow 3$ // escalated, not indicted-euro
  else if *type[d]=2 and ind[d]=1* **then**
    stratum[d]$\leftarrow 4$ // escalated, not indicted-ind
  else if *type[d]=3 and ind[d]=0* **then**
    stratum[d]$\leftarrow 5$ // indicted-euro
  else
    stratum[d]$\leftarrow 6$ // indicted-ind

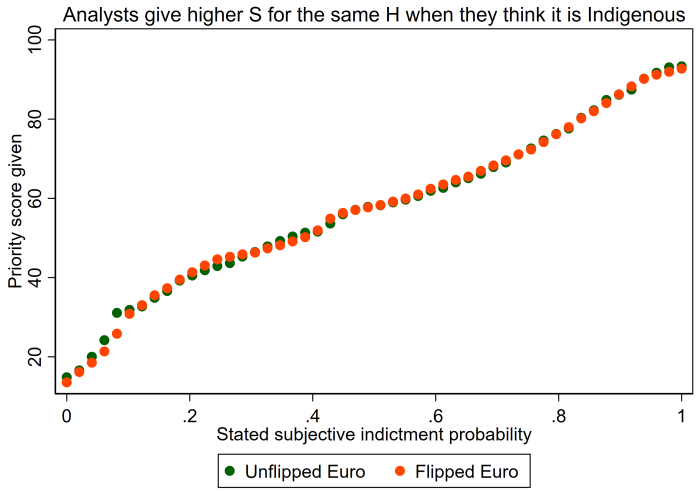
  for $a \leftarrow 1$ **to** $|A|$ **do**
    $S_a \leftarrow \{\, d \in D : \texttt{exclude}_a(d) = 0\}$ // complaints not excluded
    foreach $d \in S_a$ **do**
      $U_a^{(1)}(d) \sim \text{Uniform}(0, 1)$
    for $s \leftarrow 1$ **to** *6* **do**
      $R_{a,s} \leftarrow \{d \in S_a : \text{stratum}[d] = s\}$
      Sort $R_{a,s}$ by $U_a^{(1)}(d)$ and assign ranks $\text{caserank}_a^{(1)}(d)$
      for $i \leftarrow 1$ **to** $|R_{a,s}|$ **do**
        Let $d \leftarrow R_{a,s}[i]$
        `// Number of denuncias from each stratum assigned to each block`
        if $s \in \{1, 2, 3\}$ **then**
          $\text{block}_a(d) \leftarrow \lceil i/8 \rceil$
        else if $s = 4$ **then**
          $\text{block}_a(d) \leftarrow i$
        else if $s = 5$ **then**
          $\text{block}_a(d) \leftarrow \lceil i/4 \rceil$
        else
          $\text{block}_a(d) \leftarrow \lceil i/2 \rceil$
    foreach $d \in S_a$ **do**
      $U_a^{(2)}(d) \sim \text{Uniform}(0, 1)$
      $\text{key}_a(d) \leftarrow 10 \cdot \text{block}_a(d) + U_a^{(2)}(d)$
    Sort S_a by $\text{key}_a(d)$, then set $\texttt{showing_order}_a(d)$ accordingly

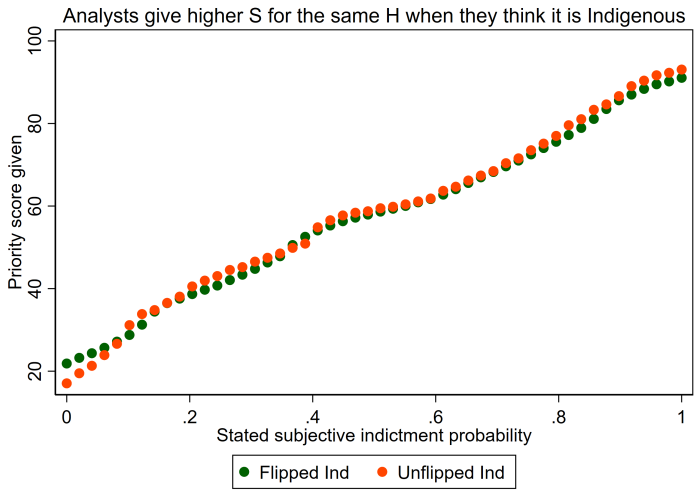
```
// 2.  Name switching within each block
for a ← 1 to |A| do
    foreach d ∈ S_a do
        U_a^(3)(d) ~ Uniform(0, 1)
    Group S_a by (block_a(d), ind[d]), rank within each group as rank_switch_a(d)
    foreach d ∈ S_a do
        artificial_denuncia_a(d) ← 0
        if block_a(d) = 1 then
            if ind[d] = 1 and rank_switch_a(d) ≤ 3 then
                artificial_denuncia_a(d) ← 1
            if ind[d] = 0 and rank_switch_a(d) ≤ 4 then
                artificial_denuncia_a(d) ← 1
        else if block_a(d) = 2 then
            if rank_switch_a(d) ≤ 4 then
                artificial_denuncia_a(d) ← 1
        else if block_a(d) = 3 then
            if ind[d] = 1 and rank_switch_a(d) ≤ 4 then
                artificial_denuncia_a(d) ← 1
            if ind[d] = 0 and rank_switch_a(d) ≤ 3 then
                artificial_denuncia_a(d) ← 1
        else if block_a(d) = 4 then
            if rank_switch_a(d) ≤ 4 then
                artificial_denuncia_a(d) ← 1
        else
            if U_a^(3)(d) ≤ 0.2419 then
                artificial_denuncia_a(d) ← 1
```

H Additional tables and figures

H.1 Prima-facie evidence for taste-based bias


Analysts give higher S for the same H to Indigenous cases
Priority score given
20
40
60
80
100
0
.2
.4
.6
.8
1
Stated subjective indictment probability
Euro
Ind


Analysts give higher S for the same H when they think it is Indigenous
Priority score given
20
40
60
80
100
0
.2
.4
.6
.8
1
Stated subjective indictment probability
Unflipped Euro
Flipped Euro


Analysts give higher S for the same H when they think it is Indigenous
Priority score given
20
40
60
80
100
0
.2
.4
.6
.8
1
Stated subjective indictment probability
Flipped Ind
Unflipped Ind

I Single-decision bias within multi-phase systems

Consider the case of the lab experiment decision to assign priority scores to a given case. We can situate our notion of bias introduced above within the many stages of a report in the CGR by explicitly accounting for bias that may occur in the arrival of the associated true report to the CGR, as well as bias in the actual outcome that is observed for that report.

Recall the decomposition of total belief-equivalent bias TOT_{jx} into TBB and BBB:

$$\underbrace{\frac{1}{2}\left\{\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x)\} + \{h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\}\right\}}_{\text{taste-based bias}} + \underbrace{h_j(i,x) + h_j(e,x)}_{\text{belief-based bias}}$$

Let us denote by $A = 1$ the event that a given a "potential case" results in an actual case or a *report*. In this case, we say that the case is "accused" (A=1). Let Y^* denote that a case would be indicted were it escalated, i.e. it is "inidictable". Let Y indicate observed indictment status of the case, so that $Y = Y^*D$ where D indicates whether the case was actually escalated. Finally, let G^* indicate whether the individual accused in a given case was actually guilty. Note that this is latent, and there will generally exist both guilty potential cases that never get accused ($A = 0$ and $G^* = 1$), as well as innocent cases that get accused and perhaps even indicted ($A = D = Y^* = 1$ but $G^* = 0$ in the latter case).

Define $\tilde{f}^*(r,x) = \mathbb{E}[G^*|R = r, X = x]$ to be the true rate of guilt among potential cases with observables r, x, and let $f^*(r,x) = \mathbb{E}[G^*|R = r, X = x, A = 1]$ be same among reports. For a given r and x, the gap $f^*(r,x) - \tilde{f}^*(r,x)$ measures selection into accusation. Let $h^*(r,x) = \mathbb{E}[Y^*|R = r, X = x, A = 1]$ be the rate of indictability among accused cases with observables r, x. The gap $h^*(r,x) - f^*(r,x)$ measures *over-indictment* for cases with observables r, x.

Define $d(r,x) := P(D = 1|R = r, X = x, A = 1)$ to be the observable escalation probability for a report with r, x, and $\tilde{h}(r,x) := \mathbb{E}[Y^*|R = r, X = x, D = 1]$ (note that $D = 1$ implies $A = 1$ so we can leave that in the conditioning event). Note that in the absence of multiple analysts that can disagree about whether to escalate a case, $d(r,x)$ would be equal to zero or to one for each r, x assuming analysts make deterministic choices given case observables. If $d(r,x) = 0$, we let $\mathbb{E}[Y^*|R = r, X = x, D = 1] := 0$ so that it remains well-defined. Let

$$h(r,x) = \mathbb{E}[Y|R = r, X = x, A = 1] = \mathbb{E}[Y^*D|R = r, X = x, A = 1] = d(r,x) \cdot \tilde{h}(r,x)$$

denote the observable indictment probability among reports with observables r, x. Consider the gap $h(r,x) - h^*(r,x) = (h(r,x) - \tilde{h}(r,x)) + (\tilde{h}(r,x) - h^*(r,x))$. The first term measures how much more likely escalated reports are to be guilty compared with unescalated reports having r, x. This term will tend to be zero if analyst esclation preferences are uncorrelated with case observables. This term will be zero in the limit of no analyst heterogeneity to lead to $d(r,x) \in (0,1)$.

Collecting, we have that

$$h(r,x) = \underbrace{\tilde{f}^*(r,x)}_{\text{true guilt probability}} + \underbrace{f^*(r,x) - \tilde{f}^*(r,x)}_{\text{accusation selection}} + \underbrace{h^*(r,x) - f^*(r,x)}_{\text{overindictment}} + \underbrace{h(r,x) - h^*(r,x)}_{\text{escalation analyst heterogeneity}} \tag{I.1}$$

With these definitions, we can write:

$$\begin{aligned} TOT_{jx} &= \underbrace{\frac{1}{2}\left\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x) + h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\right\}}_{\text{taste-based bias}} + \underbrace{h_j(i,x) + h_j(e,x)}_{\text{belief-based bias}} \\ &= \underbrace{\frac{1}{2}\left\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x) + h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\right\}}_{\text{taste-based bias}} \\ &\quad + \underbrace{\{h_j(i,x) - h(i,x)\} - \{h_j(e,x) - h(e,x)\}}_{\text{differentially-incorrect indictment beliefs}} + \underbrace{h(i,x) - h(e,x)}_{\text{accurate statistical discrimination}} \end{aligned} \tag{I.2}$$

where using Eq. (I.1) we can unpack the ASD term to write:

$$\begin{aligned} TOT_{jx} = &\underbrace{\frac{1}{2}\left\{\phi_{jev}^{-1}(\phi_{jiv}(h_j(i,x))) - h_j(i,x) + h_j(e,x) - \phi_{jiv}^{-1}(\phi_{jev}(h_j(e,x)))\right\}}_{\text{taste-based bias}} \\ &+ \underbrace{\{h_j(i,x) - h(i,x)\} - \{h_j(e,x) - h(e,x)\}}_{\text{differentially-incorrect indictment beliefs}} + \underbrace{\tilde{f}^*(i,x) - \tilde{f}^*(e,x)}_{\text{true difference in rate of guilt}} \\ &+ \underbrace{\{f^*(i,x) - \tilde{f}^*(i,x)\} - \{f^*(i,x) - \tilde{f}^*(i,x)\}}_{\text{differential accusation selection}} \\ &+ \underbrace{\{h^*(i,x) - f^*(i,x)\} - \{h^*(i,x) - f^*(i,x)\}}_{\text{differential overindictment}} \\ &+ \underbrace{\{h(i,x) - h^*(i,x)\} - \{h(i,x) - h^*(i,x)\}}_{\text{differential escalation analyst heterogeneity effect}} \end{aligned}$$

Appendix B.1 shows that in principle the sign of differential accusation selection can be identified if the econometrician observes all of the variables x that are seen by the analyst, by comparing the distribution of X among reports to that in the larger population of potential cases.

J Minimal working example of econometric framework

Suppose that case observables are $X = (X_1, X_2)$, where X_1 is the length of the denuncia (e.g. the word count) and X_2 is a binary variable indicating that the case indicates favoritism (as opposed to e.g. poor execution of public works).

We suppose that both race and monetary amounts are predictive of indictment, so that

$$P(Y = 1|R = r, X = (x_1, x_2)) = h(r,x) = \lambda_r + \rho x_1,$$

where $\rho > 0$ indicates that longer denuncias tend to be more substantive and more likely to result in an indictment.

For simplicity, we suppose that case typology X_2 does not also predict indictments. Note that we might have $\lambda_i > \lambda_e$ even if Indigenous and European officials with denuncias of the same length are equally likely to actually be guilty, but there is discrimination against Indigenous in the indictment phase.

There is a single analyst J, whose actions are characterized by the function

$$\phi_{rx_2}(h) = h + \beta \cdot \mathbb{1}(r = i) + \gamma \cdot x_2$$

If $\beta = 0$, there is no taste-based bias, and if $\gamma = 0$ the agent satisfies ERM with no V. If $\gamma \neq 0$, then the analyst satisfies ERM with $V = X_2$.

To distinguish it from $h(r, x)$, let $h_0(r, x)$ denote the single analyst's beliefs about indictment probabilities, which are off from the true indictment probability function $h(r, x)$ by an additive term τ_r:

$$h_0(r, x) = h(r, x) + \tau_r$$

In this example, we have:

- $TBB = \beta$
- $BBB = (\lambda_i - \lambda_e) + (\tau_i - \tau_e)$
- $ASD = \lambda_i - \lambda_e$
- $W_r^T = 0, \quad W_r^{\mathcal{B}} = \tau_e - \tau_i, \quad W_e^{TB} = \tau_i - \tau_e$

In this setting, using Y rather than H to implement a decomposition yields no researcher error in the estimation of total bias, and the estimate of BBB actually returns ASD only, missing the $(\tau_i - \tau_e)$ term. This bias causes TBD to in turn be biased by this same amount.

The decomposition using H works provided that the researcher controls for $V = X_2$ in estimation. In this case by Lemma 1

$$g_v^h(r, s) := g_v^h(r, s) := \mathbb{E}[H|R = r, S = s, X_2 = v] = \phi_{rv}^{-1}(s) = s - \beta \cdot \mathbb{1}(r = i) - \gamma v$$

and thus for any s: $g_v^h(i, s) - g_v^h(e, s) = \beta = TBB$. That is, the curves $g_v^h(i, s)$ will be parallel to that of $g_v^h(i, s)$, and the gap of β identifies TBB.

Now suppose that the researcher does not control for $V = X_2$, and estimate $g^h(r, s) :=$

$\mathbb{E}[H|R=r,S=s]$ rather than $g_v^h(r,s)$. Some algebra shows that

$$\begin{aligned}
g^h(r,s) &= \mathbb{E}[h_0(r,X)|R=r,\phi_{r,X_2}(h_0(r,X))=s] \\
&= \mathbb{E}[h_0(r,X)|R=r,h_0(r,X)=\phi_{r,X_2}^{-1}(s)] \\
&= \mathbb{E}[\phi_{r,X_2}^{-1}(s)|R=r,h_0(r,X)=\phi_{r,X_2}^{-1}(s)] \\
&= \mathbb{E}[h_0(r,X)|R=r,h_0(r,X)=s-\beta\cdot\mathbb{1}(r=i)-\gamma X_2] \\
&= s-\beta\cdot\mathbb{1}(r=i)-\gamma\mathbb{E}[X_2|R=r,h_0(r,X)+\beta\cdot\mathbb{1}(r=i)+\gamma X_2=s] \\
&= s-\beta\cdot\mathbb{1}(r=i)-\mathbb{E}[\gamma X_2|R=r,X_2=(s-\lambda_r-\rho X_1-\tau_r-\beta\cdot\mathbb{1}(r=i))]
\end{aligned}$$

Suppose that R, X_1, and X_2 are mutually independent, with $X_1\sim N(\mu,\sigma^2)$, $P(R=e)=p_e$, and X_2 a Bernoulli variable with $P(X_2=1)=p_x$.

Then $Z_r := (s-\lambda_r-\rho X_1-\tau_r-\beta\cdot\mathbb{1}(r=i))$ is a normally distributed random variable with mean $(s-\lambda_r-\rho\mu-\tau_r-\beta\cdot\mathbb{1}(r=i))$ and variance $\rho^2\sigma^2$. Since X_1 only takes on values 0 and γ, the above reduces to

$$\begin{aligned}
g^h(r,s) &= s-\beta\cdot\mathbb{1}(r=i)-\mathbb{E}[\gamma X_2|R=r,X_2=Z_r] \\
&= s-\beta\cdot\mathbb{1}(r=i)-\frac{f_{Z_r}(0)\cdot 0+f_{Z_r}(\gamma)\cdot\gamma}{f_{Z_r}(0)+f_{Z_r}(\gamma)} = s-\beta\cdot\mathbb{1}(r=i)-\gamma\frac{f_{Z_r}(\gamma)}{f_{Z_r}(0)+f_{Z_r}(\gamma)} \\
&= s-\beta\cdot\mathbb{1}(r=i)-\gamma\frac{\phi\left(\frac{\gamma-s+\lambda_r+\rho\mu+\tau_r+\beta\cdot\mathbb{1}(r=i)}{\rho\sigma}\right)}{\phi\left(\frac{-s+\lambda_r+\rho\mu+\tau_r+\beta\cdot\mathbb{1}(r=i)}{\rho\sigma}\right)+\phi\left(\frac{\gamma-s+\lambda_r+\rho\mu+\tau_r+\beta\cdot\mathbb{1}(r=i)}{\rho\sigma}\right)}
\end{aligned}$$

where ϕ is the standard normal probability density function. The MOT comparison then gives

$$\begin{aligned}
g^h(i,s)-g^h(e,s) = \beta+\gamma\Bigg\{&\frac{\phi\left(\frac{\gamma-s+\lambda_e+\rho\mu+\tau_e}{\rho\sigma}\right)}{\phi\left(\frac{-s+\lambda_e+\rho\mu+\tau_e}{\rho\sigma}\right)+\phi\left(\frac{\gamma-s+\lambda_e+\rho\mu+\tau_e}{\rho\sigma}\right)} \\
&-\frac{\phi\left(\frac{\gamma-s+\lambda_i+\rho\mu+\tau_i+\beta}{\rho\sigma}\right)}{\phi\left(\frac{-s+\lambda_i+\rho\mu+\tau_i+\beta}{\rho\sigma}\right)+\phi\left(\frac{\gamma-s+\lambda_i+\rho\mu+\tau_i}{\rho\sigma}\right)}\Bigg\}
\end{aligned}$$

This differs from β by a second term whose sign is not obvious. Even in the special case that $\tau_r=0$ so there is not researcher error in beliefs and $\lambda_r=0$ so there is no accurate statistical discrimination, the above simplifies to

$$g^h(i,s)-g^h(e,s)=\beta+\gamma\left\{\frac{\phi\left(\frac{\gamma-s+\rho\mu}{\rho\sigma}\right)}{\phi\left(\frac{-s+\rho\mu}{\rho\sigma}\right)+\phi\left(\frac{\gamma-s+\rho\mu}{\rho\sigma}\right)}-\frac{\phi\left(\frac{\gamma-s+\rho\mu+\beta}{\rho\sigma}\right)}{\phi\left(\frac{-s+\rho\mu+\beta}{\rho\sigma}\right)+\phi\left(\frac{\gamma-s+\rho\mu+\beta}{\rho\sigma}\right)}\right\}$$

With $\beta \neq 0$, the second term only vanishes in the knife-edge case that

$$\frac{\phi\left(\frac{-s+\rho\mu+\beta}{\rho\sigma}\right)}{\phi\left(\frac{\gamma-s+\rho\mu+\beta}{\rho\sigma}\right)} = \frac{\phi\left(\frac{-s+\rho\mu}{\rho\sigma}\right)}{\phi\left(\frac{\gamma-s+\rho\mu}{\rho\sigma}\right)},$$

which cannot hold for all s.

K Further appendices to be organized in appendices

K.1 Heterogeneity in the components of bias

Consider any set of covariates $V \subset X$, which may be larger than is necessary to satisfy ERM, and a set of analysts $\mathcal{J}$. Theorem 1 naturally extends to decompose TOT into BBB and TBD within the pairing of a value v of V and the set $\mathcal{J}$ of analysts. For example:

$$\mathbb{E}\left\{\mathbb{E}[g^h_{JV}(i,S)|e,J,B,V=v,F=1] - \mathbb{E}[g^h_{JV}(e,S)|e,J,B,V=v,F=0] \mid R=e, J\in\mathcal{J}, V=v\right\}$$

yields $\mathbb{E}[h(i,X) - h(e,X)|R=e, J\in\mathcal{J}, V=v]$, while

$$\mathbb{E}\left\{\mathbb{E}[g^h_{JV}(e,S)|e,J,B,V=v,F=1] - \mathbb{E}[g^h_{JV}(i,S)|e,J,B,V=v,F=1] \mid R=e, J\in\mathcal{J}, V=v\right\}$$

yields $\mathbb{E}[\phi^{-1}_{Jev}(\phi_{Jiv}(h(i,X))) - h(i,X)|R=e, J\in\mathcal{J}, B, V=v]$. As for total bias,

$$\mathbb{E}\left\{\mathbb{E}[g^h_{JV}(e,S)|e,J,B,V=v,F=1] - \mathbb{E}[g^h_{JV}(e,S)|e,J,B,V=v,F=0] \mid R=e, J\in\mathcal{J}, V=v\right\}$$

yields $\mathbb{E}[\phi^{-1}_{Jev}(\phi_{Jiv}(h(i,X))) - h(e,X)|R=e, J\in\mathcal{J}, B, V=v]$, and similarly for Indigenous cases.