Cognitive Distortions in Policing Decisions

Juan B. González, Santiago De Martini
& Santiago Perez-Vincent

University of Southern California & IADB

Motivation

Stylized Fact: police officers exhibit significant racial bias.

Black and Hispanic people:

- Are stopped and searched more often (Pierson et al. 2020)
- Are exposed to more policing in their neighborhoods (Vomfell and Stewart 2021)
- Are arrested and incarcerated more often (Chen et al. 2023)
- Suffer more police brutality (Fryer 2019) and more police murders (Edwards, Lee, and Esposito 2019).

Police Decision Making

How does the police make decisions?

- Objective is maximizing arrests, not minimizing crime (Stashko 2023).
- Need a prediction of where crimes happen (where to patrol) and who commits them (who to stop, search, profile...)
- Analyze previous crime data in order to make these predictions.
- Predict racialized people to be more likely to commit a crime: **statistical discrimination**.

These beliefs are often **inaccurate**: conditional on being stopped, racialized people are less likely to have committed a crime (Pierson et al. 2020; Vomfell and Stewart 2021).

Discrimination

Discrimination can be separated into 2 categories (Bohren, Imas, and Rosenberg 2019):

Taste-based:
Agent receives utility from increasing differences between groups.

Beliefs-based:

- Accurate statistical discrimination: differences in beliefs follow actual differences in distributions.
- Inaccurate statistical discrimination: differences in beliefs don't correspond to actual differences in distributions.

Inaccurate statistical discrimination

Stereotyping/Representativeness:

- Exaggerate actual differences in distributions across groups.

Inaccurate statistical discrimination

Stereotyping/Representativeness:

- Exaggerate actual differences in distributions across groups.

Data Selection:

- Over-policing in Black and Hispanic neighborhoods leads to more likelihood of catching crimes happening there.
- This leads to higher arrest rates in these neighborhoods.
- Police decision makers see the higher arrests rates and infer higher criminality in these areas, neglecting the selection of data.
- Areas perceived to be more criminal are subject to even more policing, creating a vicious loop.
- Over-policing explains around 60% of the racial gap in arrests (Chen et al. 2023).

Case: Los Angeles Police Department

Between 2011-2020, LAPD used the PredPol algorithm to predict where crime would happen.

- The algorithm used previous crime data to predict the location of future crimes.
- As there were more policing and more arrests in racialized neighborhoods, crimes were predicted there.
- LAPD would overpatrol these areas, stop and arrest people there, and feed this data into the algorithm.
- Then the algorithm would predict even more crime in these areas, creating a vicious loop.

LAPD has a budget of $3.4 billions (2024-2025).

Causes of statistical discrimination

Police officers might be making statistical inference based on selected data (**selection neglect**) and potentially distorted perceptions (**representativeness**).

This paper

- Models how selection neglect and representativeness distort belief updating and decision making.
- Designs an experimental setting to test and estimate the model.
- Relates individual level estimates with discrimination outside the lab.
- Tests interventions to reduce discrimination in police decision making.

Model

A model on selection neglect and representativeness

Inspired by theoretical and experimental work on:

- Selection neglect by Enke (2020), Barron, Huck, and Jehiel (2024), and Hübert and Little (2023).
- Stereotypes and representativeness by Bordalo et al. (2016) and Esponda, Oprea, and Yuksel (2023).
- Discrimination by Bohren, Imas, and Rosenberg (2019) and Campos-Mercade and Mengel (2024).

The setting

- A decision maker (police officer) must estimate the type t of some individual (criminality).
- The individual belongs to a group g (e.g., gender, race, neighborhood), observed by the DM.
- Types are drawn from the distribution $f(t|g)$, so $t \sim \mathcal{N}(\mu_g, \sigma^2)$.
- The DM observes a noisy signal s, e.g., previous arrests or reports.
- Signals are drawn from the distribution $h(s|t,g)$, so $s \sim \mathcal{N}(t, \upsilon_g^2)$.
- Signals are then $s = t + \epsilon_g$, where $\epsilon_g \sim \mathcal{N}(0, \upsilon_g^2)$.

Optimal updating

A Bayesian DM would update their type prediction $\hat{t}$ following:

$$\hat{t} = \omega_g \mu_g + (1 - \omega_g)s \tag{1}$$

where ω_g is the weight put on the group prior, with the optimal Bayesian weight being the relative precision of the prior wrt the signal: $\omega_g^* = \frac{v_g^2}{\sigma^2 + v_g^2}$.

Representative Prior Distortion

Representative type: a type that is more likely to be observed in one group than in another. Bordalo et al. (2023) model representativeness as $R(t, g, -g) := \frac{f(t|g)}{f(t|-g)}$.

Representative Prior Distortion

Representative type: a type that is more likely to be observed in one group than in another. Bordalo et al. (2023) model representativeness as $R(t, g, -g) := \frac{f(t|g)}{f(t|-g)}$.

Representative types are more easily recalled, distorting the perception of the prior:

$$\tilde{f}(t|g) = \kappa f(t|g) R(t, g, -g)^{\gamma^p} \tag{2}$$

where κ is a normalization factor and γ^p captures how prone is the agent to distorting priors by representativeness.

Representative Prior Distortion

Under RPD, the DM updates using the distorted prior distribution $\tilde{f}(t|g)$, centered around a distorted mean $\tilde{\mu}_g = \mu_g + \gamma^p(\mu_g - \mu_{-g})$.

The optimal prediction becomes:

$$\begin{aligned} \hat{t} =& \omega_g \tilde{\mu}_g + (1 - \omega_g)s = \\ & \omega_g \mu_g + \underbrace{\omega_g \gamma^p(\mu_g - \mu_{-g})}_{\Delta^{RPD}} + (1 - \omega_g)t + \epsilon_g \end{aligned} \tag{3}$$

Selection Neglect

- Let $p_g \in [0, 1]$ be the level of policing over group g, and t the criminality of an individual/area.
- The data (observed number of reports or arrests) is selected depending on the level of policing: $p_g t$. The data represents the true type only if there is full surveillance of an area or an individual is stopped.
- If selection is accounted for, the DM discounts selection and considers (correctly) the signal to be $s = t + \epsilon_g$.
- If selection is neglected, the selected data is taken directly as a signal of criminality $s = p_g t + \epsilon_g$.

Selection Neglect

The perceived signal $\tilde{s}$ is a convex combination of the incorrect $(p_g t)$ and correct signal (t):

$$\tilde{s} = (p_g t)^{\lambda} t^{1-\lambda} + \epsilon_g = p_g^{\lambda} t + \epsilon_g \tag{4}$$

where λ captures the degree of selection neglect. The DM is unaware of their level of selection neglect, so they still believe their perceived signal to be drawn from $h(s|t, g)$.

Selection Neglect

The optimal prediction becomes:

$$\begin{aligned}\hat{t} =& \omega_g \mu_g + (1-\omega_g)\tilde{s} = \\ & \omega_g \mu_g + (1-\omega_g) \underbrace{p_g^{\lambda}}_{\Delta^{SN}} t + \epsilon_g \end{aligned} \tag{5}$$

RPD and Selection Neglect

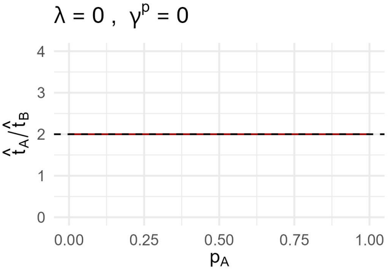
The optimal prediction becomes:

$$\begin{aligned}\hat{t} =& \omega_g \tilde{\mu}_g + (1-\omega_g)\tilde{s} = \\ & \omega_g \mu_g + \underbrace{\omega_g \gamma^p (\mu_g - \mu_{-g})}_{\Delta^{RPD}} + (1-\omega_g) \underbrace{p_g^\lambda}_{\Delta^{SN}} t + \epsilon_g \end{aligned} \tag{6}$$

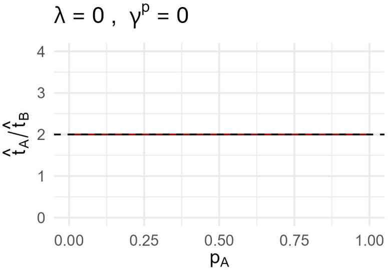
An example with RPD and Selection Neglect

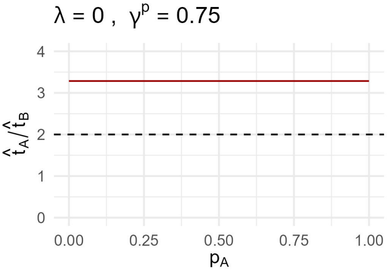
- Let $\mu_A = 30, \mu_B = 15$.
- Let the type drawn be the mean of each group: $t_g = \mu_g$.
- Thus, the true $\frac{t_A}{t_B} = 2$
- Assume $p_A + p_B = 1$
- Let $\omega_A = \omega_B = 0.5$.
- How does the prediction depend on the level of policing (p_A), of RPD (γ^p) and selection neglect (λ)?

An example with RPD and Selection Neglect

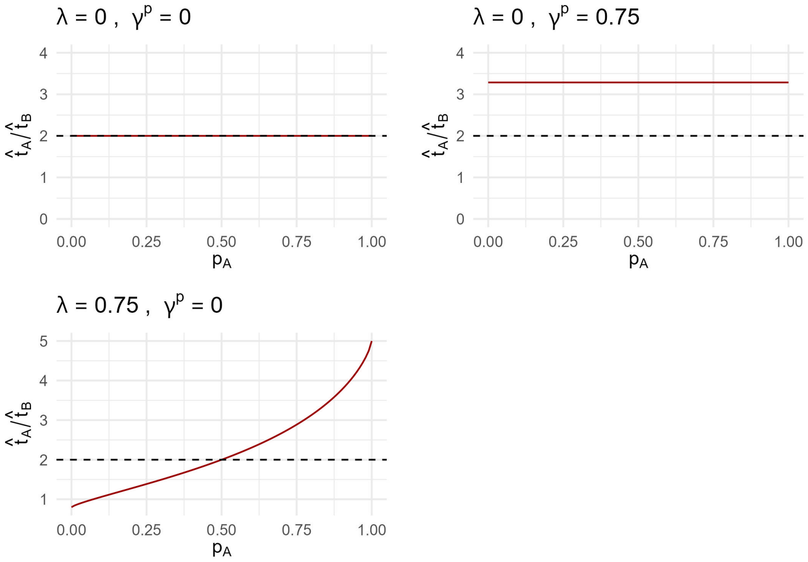


An example with RPD and Selection Neglect

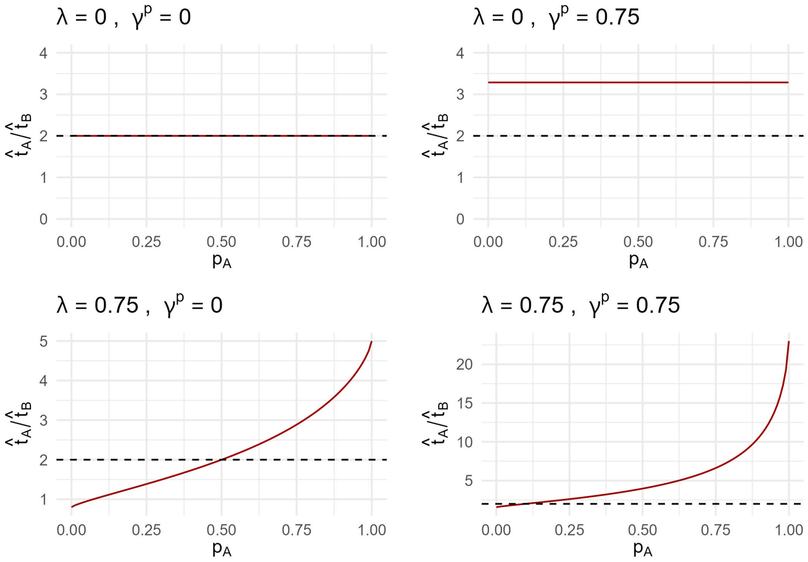




An example with RPD and Selection Neglect



An example with RPD and Selection Neglect



Dynamics

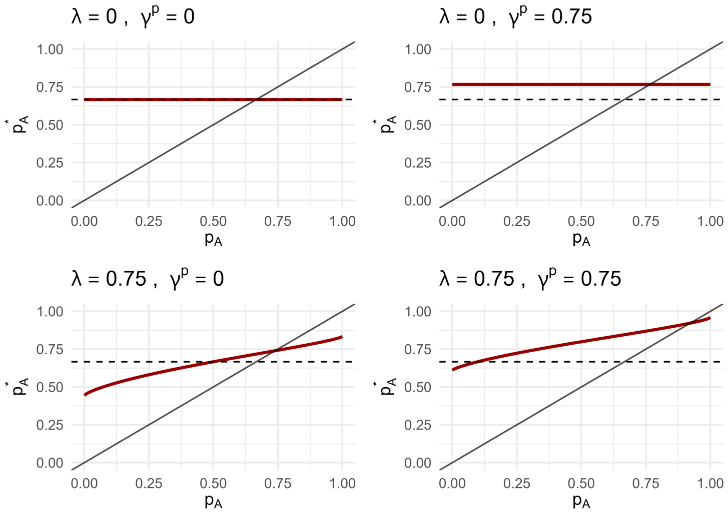
Let the decision of how much to patrol neighborhood A be given by its relative predicted criminality wrt to neighborhood B, or the probability of stopping individual A by their relative predicted criminality:

$$p_A^* = \frac{\hat{t}_A}{\hat{t}_A + \hat{t}_B}$$

Notice that predicted types depend on the previous level of policing p_A, and so does the decision. Thus we have a Best Response function $p_A^*(p_A)$.

Using the same example as before, the true $p_A^* = \frac{30}{30+15} = 0.667$

Dynamics



Best Response Function $p_A^*(p_a)$ for different values of RPD (γ^p) and Selection Neglect (λ)

Model Summary

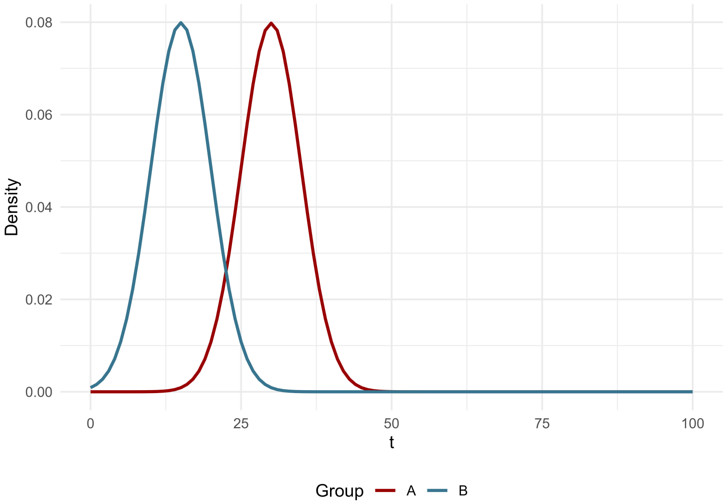
- Representativeness can distort the perception of the prior, exaggerating real differences in distributions across groups.
- Selection neglect can create dynamics of distorted perception leading to inaccurate decision making.
- Both can combine to generate inaccurate statistical discrimination.

Experimental Design

Experiment 1

- Two demographic groups, the Reds and the Blues.
- Each neighborhood (a continuum of them) belongs to one of the two groups.
- Each neighborhood has a different level of criminality (type t), that we can take as % criminal in each area.
- Type t is drawn from the distribution $f(t|g)$ (prior).

Experiment 1



Prior type distribution $f(t|g)$

Experiment 1

- Police patrols are sent to each neighborhood, and they report a signal s of their criminality t.
- The signal is centered around the observed criminality but it's noisy (sometimes patrols overestimate, sometimes they underestimate).
- Each neighborhood has a different number of police patrols that determine the level of policing p_g.
- How much of the crime actually happening is observed depends on the level of policing p_g.
- The reported level of criminality is thus $s = p_g t + \epsilon$.
- Priors of the distribution of criminality across groups are shown at the beginning.

Experiment 1

- Both priors and signals are noisy.
- The relative noise of each should determine the weight of the prior when updating (ω_g).
- Both dispersions are provided during the experiment.

Experiment 1

Information:

Neighborhood	**A**	**B**
Group	Blue	Red
N of Patrols (/10=p_g)	6	4
Noisy Report s	11	10
Criminality t	15	30
Accurate Report $p_g t$	9	12

Decision: What's your prediction of criminality for each neighborhood? $(\hat{t}_A, \hat{t}_B)$

Dynamics: In next round, allocation of patrols will follow the previous prediction $\frac{\hat{t}_A}{\hat{t}_A+\hat{t}_B}$.

Experiment 1

Treatments:

- **Benchmark:** No selection (N of patrols always fixed at 1:1), no representativeness ($\mu_{red} = \mu_{blue}$).
- **Selection:** Dynamic selection, no representativeness.
- **RPD:** No selection, introduce representativeness ($\mu_{red} >> \mu_{blue}$).
- **SN & RPD:** Dynamic selection and representativeness.

Experiment 2

Goal: Measure statistical discrimination against racial groups, at the individual level, using a simple and **short** task.

Ideas:

- Direct elicitation of predicted criminality of suspects of varying races, joint with second-order elicitation?
- Conjunction fallacy — *what's more likely, to be of X race, to be a criminal or to be of X race & criminal?*
- Reduced version of IAT?
- Other?

Questions

- What interventions could be effective to reduce these drivers of statistical discrimination?
- What could we do to measure taste-based discrimination?
- What external validity tests could be useful?
- Any experience working with some US Police Department?