How do disparities reproduce themselves? "Ground truth" inference from utility-maximizing agent's sampling behavior

Yuan Meng
yuan_meng@berkeley.edu
Department of Psychology
University of California, Berkeley

Fei Xu
fei_xu@berkeley.edu
Department of Psychology
University of California, Berkeley

Abstract

In an ideal world, evidence of disparities motivates people to fight them; in reality, disparities often "reproduce" themselves. Upon seeing the police stopping some groups at higher rates, people may believe members of these groups are more prone to crime and therefore seek more punitive measures against them. In this paper, we argue that even without stereotypic associations linking some groups with crime, people may still reproduce observed disparities via rational inference: Assuming an agent is knowledgeable about a target trait's "hit rates" in different groups and acts to maximize the expected utility of checking, you may infer that groups checked more often have higher hit rates. In Experiment 1, this "Naïve Utility Calculus" successfully captured how people inferred the hit rate in a population based on how often an agent sampled from it ("check rate"). In Experiment 2, when hit rates in the samples were revealed, people predominantly relied on this new information more heavily than the agent's check rates. Our work both provided a novel explanation for why people reproduce disparities and a potential intervention to combat such a tendency.

Keywords: social cognition; disparities; theory of mind; computational modeling

Introduction

In Oakland, California, around 60% of the police stops between 2013 and 2014 were of African Americans, who only make up about 28% of the local population and were no more likely than members of other races to carry contraband (Hetey, Monin, Maitreyi, & Eberhardt, 2016). Not only does intense proactive policing tend to induce psychological stress and future crimes in the African American community (Del Toro et al., 2019), it's also the first step towards disparities in areas with more dire consequences, such as arrests and incarceration. As scientists and educators, what can we do to help combat racial disparities in proactive policing?

Perhaps if people saw what we saw, they would be eager to end racial disparities. However, numbers don't always speak for themselves: Evidence of disparities often leads people to support the very system that has created them in the first place (Hetey & Eberhardt, 2018). For instance, when presented with a "Blacker" prison compared to a "less Black" one (45% *vs.* 25% inmates were African Americans), White voters became more supportive of laws that severely punish repeated offenders (Hetey & Eberhardt, 2014). Why would people support laws that may intensify racial disparities after being made more aware of their existence? Hetey and Eberhardt (2014, 2018) argued that this paradox could be explained by stereotypic associations linking Blacks with crime: A Blacker prison made people more fearful of crime and the fear prompted them to support stricter criminal law.

Are pre-existing stereotypes such as the "Black-crime association" necessary for observers to reproduce disparities? Imagine, as a new quality inspector at a factory, you shadowed the most experienced inspector on your first day there. They checked 7 out of every 10 X products and only 3 out every 10 Y products. Inspecting a product incurs a small cost every time but catching a defective one is far more rewarding. When it's your turn, would you check X or Y more often? If your answer is "X", is it because you have stereotypic associations linking it with low quality? This explanation is unlikely since you've only just seen these products. Alternatively, your preference may be a result of *rational inference*: Assuming the inspector is knowledgeable about product quality and acts to maximize the expected utility of checking (the expected reward from finding defective products minus the total cost of checking), it only makes sense if X has a higher defect rate than Y—so you should also check X more often, just like your predecessor did. This kind of *Naïve Utility Calculus* (NUC, Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016) explains how we might end up reproducing disparities we observe without having any stereotypes from the get-go.

Inferring "ground truth" from sampling processes

It's not a new idea that sampling behavior is a window into an intentional agent's preference. Even young children (Kushnir, Xu, & Wellman, 2010) and infants (Wellman, Kushnir, Xu, & Brink, 2016) understand that a person selecting rare toys at a disproportionately high rate prefers these toys and will choose the same type in the future. It's also been proposed that such inferences arrive from a *principle of efficiency* (Jara-Ettinger, Sun, Schulz, & Tenenbaum, 2018): A utility-maximizing agent wouldn't go out of their way to seek out the rarer option if they didn't like it better.

The police scenario differs from previous studies in that we care *not* about how people use sampling to infer others' preference but some "ground truth" about the world. To our best knowledge, only one study (Gweon, Tenenbaum, & Schulz, 2010) has looked at ground truth inference from non-random sampling. In their study, infants had to infer was whether a property (squeaking) found in one category (blue balls) extended to another (yellow balls). When most balls were yel-

low but the experimenter only sampled blue balls which all squeaked, infants didn't expect yellow balls to also squeak (*"If yellow balls also squeak, why didn't she show me?"*).

Ground truth inference in our study goes beyond binary: Rather than whether a trait extends to a group, people must infer the prevalence of that trait. After all, members of all races commit crime but what observers believe drives the police's sampling behavior is the magnitude of crime rates. Furthermore, infants in Gweon et al.'s (2010) study had to simultaneously infer whether the experimenter was sampling from all balls or just squeaking balls. This joint inference captures observers' uncertainty about both the police's intention and knowledge. However, before factoring in the agent's intention, we want to first establish whether people can make accurate inference about "crime rates" based on sampling. We made it clear that the agent is cost-effective and knowledgeable about true "crime rates" in different groups. In the future, it's worth exploring joint mental state/ground truth inference as a model of how people learn from a rational agent's sampling behavior and also a potential anti-bias intervention.

By viewing the reproduction of disparities from a rational perspective, by no means do we wish to "rationalize" it or downplay the role of stereotypes discussed in a vast body of literature (e.g., Eberhardt, Goff, Purdie, & Davies, 2004; Payne, 2001). Rather, what we are ultimately searching for is an effective way to fight disparities and if we find that rational inference alone has the power to reproduce disparities, we need further measures than just doing away with stereotypes.

Study overview

To capture key elements of proactive policing without evoking pre-existing racial stereotypes, we designed a novel arcade game called the "Golden Ticket" (Figure 1). In this game, a robot chicken lays one egg at a time, which is either empty or has a golden ticket that can be redeemed for a grand prize. Each egg costs a token and tokens are expensive and come in limited quantities. Players can pass on eggs that they don't want to buy. This game is designed to simulate real-life police encounters. A robot chicken and the eggs it lays are akin to a social group and members of that group and the player symbolizes the police officer. Winning a golden ticket and catching a criminal are both highly rewarding yet checking incurs a cost every time, be it a token or time and energy. Each chicken has a fixed *"hit rate"* (the probability of laying an egg that has a golden ticket) known to an experienced player Alex. Alex also doesn't spend tokens unnecessarily. Participants watched Alex play a series of robot chicken. After each, they were asked to infer the chicken's hit rate and indicate whether they wanted to buy a new egg from it.

In Experiment 1, the eggs' content was not revealed. We looked at whether participants could infer hit rates based on *"check rates"* (the probability that Alex buys an egg from a chicken) alone. Their inferences were compared to predictions generated by a Naïve Utility Calculus model. In Experiment 2, Alex opened the eggs after each round. We examined whether participants used both the hit rate and the check rate in the sample to infer the true hit rate in the population or if they relied on one source of information. We created three computational models to compare against human performance, each corresponding to an aforementioned possibility (hit rate only, check rate only, check + hit rates).

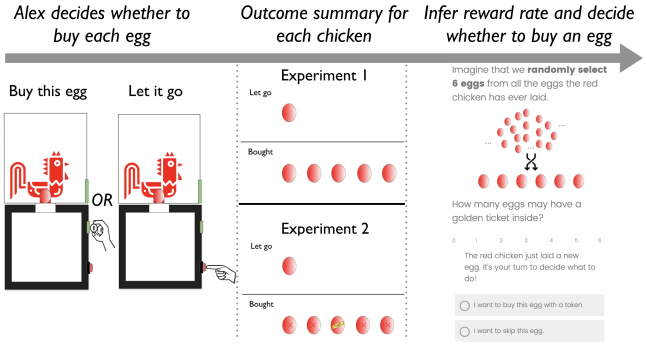

Figure 1: The Golden Ticket game: A robot chicken laid a total of 6 eggs one at a time. Alex, the best player who knew every chicken's hit rate and refrained from unnecessary purchase, decided whether to buy each egg or let it go. In Experiment 1, participants only saw how many eggs Alex bought but not the content of the eggs. In Experiment 2, Alex opened the eggs they bought, revealing which ones had tickets and which ones were empty. At the end of each trial, participants were asked to estimate the hit rate of the robot chicken and indicate whether they want to buy a new egg laid by this chicken.

Computational Modeling

Naïve Utility Calculus models

We created computational models based on the Naïve Utility Calculus (NUC) (Jara-Ettinger et al., 2016) to predict how learners infer each chicken's hit rate from Alex's check rate. The key assumption is that Alex maximizes the expected rewards of winning golden tickets relative to the total costs of buying eggs. Suppose each ticket has a value of V, each token costs C, and the true hit rate of a chicken is θ, then the expected utility of buying an egg from a given chicken is:

$$\mathrm{E}[U(buy)] = V\theta - C \tag{1}$$

The higher the expected utility $\mathrm{E}[U(buy)]$, the more likely that Alex will buy an egg, hence the higher the check rate μ. We assume that Alex's choice behavior follows a softmax choice rule (Sutton & Barto, 1998). In the case of binary choices, the softmax function becomes the logistic function:

$$\mu = \frac{1}{1+\exp(-\mathrm{E}[U(buy)])/\tau} \tag{2}$$

where τ is a temperature parameter controlling the level of stochasticity in Alex's decisions. When $\tau \to 0$, Alex will always buy an egg if $\mathrm{E}[U(buy)] > 0$; when $\tau \to \infty$, Alex just

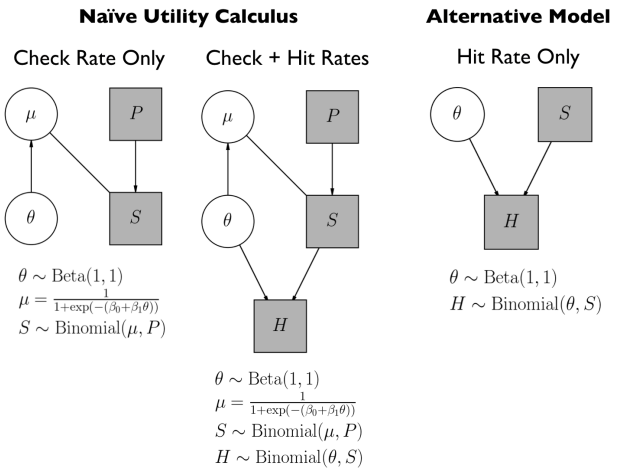

Figure 2: Computational models of hit rate inference. In the two Naïve Utility Calculus (NUC) models, the check rate μ is inferred from the number of eggs Alex bought (S) out of the number of eggs a chicken laid (P). μ is linked to the true hit rate θ via a logistic function, whose intercept β_0 and slope β_1 are estimated from participants' inferred hit rates and purchase decisions. In Experiment 1, θ can only be inferred from μ. In Experiment 2, θ can be simultaneously inferred from μ and the number of eggs with tickets (H) out of P. Alternatively, people may ignore μ and assume that the observed hit rate is the true hit rate, in which case "hit rate only" model captures their inference.

randomly decides between buying an egg and letting it go.

The three free parameters V, C, and τ can be absorbed into the slope $\beta_1 = V/\tau$ and the intercept $\beta_0 = -C/\tau$ of the logistic function in Equation 2, which then becomes:

$$\mu = \frac{1}{1+\exp(-(\beta_0+\beta_1\theta))} \tag{3}$$

We can estimate β_0 and β_1 from participants' inferred hit rates and subsequent purchase decisions. Plugging the estimates back in Equation 3 allows us to infer what participants thought the true hit rate θ was based on the observed check rate μ—assuming that they used the same utility function to interpret Alex's decisions and make their own decisions[1].

In both NUC models, μ is inferred from the number of eggs Alex bought, S, out of the number of eggs each chicken laid, P: $S \sim \text{Binomial}(\mu, P)$. In Experiment 1 with only check rate information, θ is solely inferred from μ via Equation 3. In Experiment 2, θ can be simultaneously inferred from μ and the number of eggs with tickets, H, out of the number of eggs Alex bought, S: $H \sim \text{Binomial}(\theta, S)$. If participants ignore new hit rate information, then their inferential process is captured by the same "check rate only" model in Experiment 1.

[1]This is not necessarily true. People's own utility function can differ from what they think an another agent's utility function is. For instance, Liu, McCoy, and Ullman (2019) found that most of their participants saw others as less or more risk averse than—but not the same as—themselves. However, since what we care about is each model's relative performance, this difference is not too concerning.

Alternative model

In Experiment 2 where hit rate information was also provided, participants may ignore check rates and assume that each chicken's observed hit rate is its true hit rate. In this "hit rate only" model with no utility concerns, θ is solely determined by H (the number of eggs with tickets) and S (the number of eggs Alex bought): $H \sim \text{Binomial}(\theta, S)$.

Experiment 1

Methods

Participants Sixty-two adult residents of the United States (mean age = 38.5 years) participated in Experiment 1 through Amazon Mechanical Turk. A past acceptance rate of at least 95% was required for participation. Another 31 participants were excluded for failing any of the five instruction check questions. All participants gave informed consent prior to the study and were paid $2 for about 15-20 minutes of their time.

Procedure Participants first watched a video introducing the "Golden Ticket" game and were then tested on the utility structure of this game, namely rewards (a golden ticket or none) and costs (a token or none). Next, participants learned that different robot chickens may have different hit rates and the best player Alex knows each chicken's hit rate and avoids spending tokens unnecessarily. It was also emphasized that while each chicken has a fixed hit rate *on the long run*, the hit rate in each batch of eggs is subject to random fluctuations. To ensure that participants understood all the critical information, we tested them on what makes Alex the best player (because they are knowledgeable about hit rates and cost-effective) as well as possible fluctuations in hit rates.

Those who passed both rounds of instruction checks within two attempts were allowed to continue to the main experiment, where they watched short videos of Alex playing 12 different robot chickens, each distinguished by a unique color. The total number of eggs laid by each chicken was held constant (6 eggs) while the number of eggs Alex bought varied (2 chickens: bought 1 or 6 out of 6 eggs; 4 chickens: bought 2 or 5 out of 6 eggs). The order of the 12 chickens was randomized for each participant. Once Alex made decisions for all 6 eggs laid by a chicken, participants were reminded of how many eggs they bought or let go on a summary page.

To measure hit rate inference, we asked participants to guess that out of 6 eggs randomly selected from all the eggs a chicken has ever laid, how many might have golden tickets. Finally, they were asked to decide whether they would buy a new egg from this chicken or pass on the opportunity.

Results

The central question that inspired this research is whether people would still reproduce observed disparities even without existing stereotypes. In Experiment 1, this question trans-

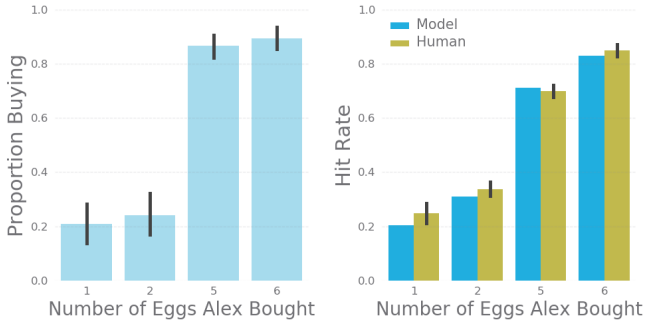

Figure 3: Results from Experiment 1: (Left) The proportion of participants deciding to buy a new egg as a function of the number of eggs Alex bought. (Right) Hit rates of robot chickens predicted by the model vs. inferred by participants. (Error bars indicate the 95% confidence intervals.)

lates to whether participants' purchase decisions were influence by Alex's check rates. From Figure 3 (left), it's obvious that far more participants chose to buy a new egg when Alex bought 5 or 6 eggs compared to only 1 or 2. To test this observation more rigorously, we fit a generalized linear mixed model (GLMM) using the R package `lme4` (Bates, Mächler, Bolker, & Walker, 2015) with participants' purchase decisions as the outcome variable, the number of eggs Alex bought as the fixed effect, as well as random intercepts and slopes for participants and trials[2]. Consistent our observation, with each additional egg Alex bought, the odds ratio for participants buying an egg increased by 2.36, which was significantly higher than chance, Wald's $\chi^2 = 14.03$, $p = .00018$.

Were participants "mindlessly" copying Alex's sampling behavior or did they use it to infer robot chickens' hit rates in a "rational" way? To answer this question, we examined whether participants' hit rate inferences can be captured by a NUC model. First of all, we estimated parameters values in Equation 3 to be $\beta_0 = -2.56$ and $\beta_1 = .98$ using the Python library `PyMC3` (Salvatier, Wiecki, & Fonnesbeck, 2016) based on participants' inferred hit rates and purchase decisions. Then we implemented a "check rate only" NUC model to predict how people should infer each chicken's hit rate from Alex's check rate and compared model predictions against participants' actual inferences. As you can see in Figure 3 (right), this NUC model captured all the qualitative trends across all trial types. Model predictions were strongly correlated with human responses, Pearson's $r = .79$, $p < .001$.

Discussions

Before playing the Golden Ticket, participants had no expectation for each robot chicken's hit rate. Yet, after watching a knowledgeable, utility-maximizing agent Alex playing the game, participants quickly followed their lead, buying more eggs from chickens that Alex was more likely to buy from. This result shows that participants ended up reproducing observed disparities in check rates without having prior stereotypes of chickens' hit rates. At least when explicitly asked to, participants could do more than just blindly copying Alex's sampling behavior: Each chicken's hit rate that they inferred from Alex's check rate closely matched the prediction generated by a simple Naïve Utility Calculus model. Perhaps through a similar inferential process, a naïve observer of police encounters may conclude that groups under heavier scrutiny have higher crime rates and will check members from these groups more often when given the opportunity.

What if groups checked more often have the same or even lower hit rates? One of the most striking and informative findings from the Oakland police stop data is that African Americans who were stopped far more often than other races were no more likely to carry contraband. If people trust a knowledgeable, utility-maximizing agent blindly, they may disregard this new hit rate information as a "fluke". It's also possible that they ignore the agent's sampling behavior and solely focus on observed hit rates. Alternatively, people may consider both the hit rates and the check rates in the sample to infer group hit rates. To examine these possibilities, we conducted Experiment 2 where Alex opened the purchased eggs to reveal which ones had tickets and which ones were empty.

Experiment 2

Methods

Participants Sixty-seven adult residents of the United States (mean age = 36.9 years) participated in Experiment 2 through Amazon Mechanical Turk. To participate, one must have a past acceptance rate of 95% or above and not have taken part in Experiment 1. Another 30 participants were excluded for failing any of the five instruction check questions. All participants gave informed consent prior to the study and were paid $2 for about 15-20 minutes of their time.

Procedure The procedure for Experiment 2 was identical to that for Experiment 1, except that Alex opened the eggs they bought in the end to reveal which ones had tickets and which ones were empty. There were 12 unique combinations of how many eggs had tickets out of how many eggs Alex bought. Of these, there were 6 critical trials where Alex bought most or all eggs but most or all were empty (0 or 1 out of 6 eggs or 0 or 1 out of 5 eggs had tickets) or few eggs but all had tickets (1 out of 1 egg or 2 out 2 eggs had tickets). In case participants suspected that Alex was misleading or not actually knowledgeable, we added 6 "filler" trials where Alex bought most or all eggs and most or all had tickets (5 or 6 out of 6 eggs or 4 or 5 out of 5 eggs had tickets) or few eggs and none had tickets (0 out of 1 egg or 0 out 2 eggs had tickets). The order of presentation was randomized across participants.

Results

After observing both the check rate and the hit rate in an egg sample, how did participants infer the chicken's true hit rate?

[2]Here we implemented `lme4` models in Python via `rpy2` to keep in the same environment. The formula for the full model was: `buy ~ checked + (1 | participant) + (1 | trial)`.

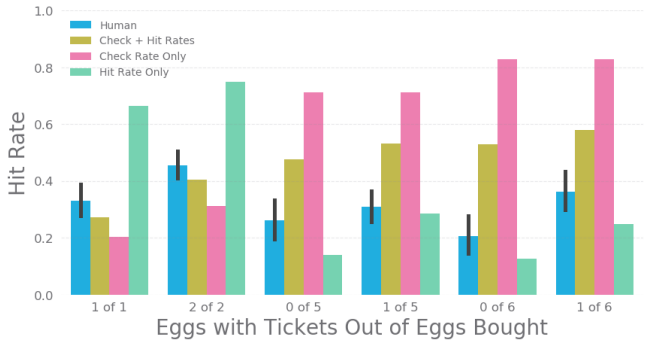

Figure 4: Results from Experiment 2: Hit rates of robot chickens predicted by three models ("check + hit rates", "check rate only", "hit rate only") vs. inferred by participants. (Error bars indicate the 95% confidence intervals.)

We created the three models to capture three hypotheses (Figure 2): 1) a NUC model inferring hit rates from both observed check rates and hit rates ("check + hit rates"), 2) another NUC model only inferring from observed check rates ("check rate only"), and 3) a non-NUC "hit rate only" model.

Based on participants' inferred hit rates and purchase decisions, we estimated the parameter values in Equation 2 to be $\beta_0 = -2.79$ and $\beta_1 = 1.10$. The we implemented the above models using the estimates (with the exception of the "hit rate only" model, which doesn't require these parameters). To see which model resembled human performance the most, we used the root mean square error[3] (RMSE) to measure the dissimilarity between model predictions and human inferences. Lower RMSE indicates better model fit. As it turned out, the "check rate only" model's predictions strayed furthest away from human inferences (RMSE = .37) whereas the "check + hit rates" model and the "hit rate only" model (RMSE = .27 and RMSE = .26, respectively) performed similarly.

Figure 4 compared model predictions with human inferences across 6 critical trials[4] where observed hit rates and check rates were at odds with each other (i.e., when one was high, then the other was low). Not a single model captured all the qualitative trends across all trial types. Among the three, the "check rate only" model was the least accurate one, grossly overestimating hit rates when Alex bought many eggs and underestimating when few. The "check + hit rates" model and the "hit rate only" model both captured part of the trends: When Alex bought just few eggs but all had tickets, the "check + hit rates" model was the most accurate; when Alex bought bought many eggs but few had tickets, the "hit rate only" model performed most similarly to humans.

[3]In this case, RMSE $= \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(\theta_{m_i} - \theta_{h_i})^2}$, where n is the number of participants and θ_m and θ_h are hit rates predicted by a model and inferred by participants, respectively.

[4]Models made similar predictions on non-critical trials. In the interest of space and model comparison, we didn't plot these trials.

Discussions

All in all, no model captured participants' hit rate inferences exactly when both check rate and hit rate information was available. Why did the "check + hit rates" model and the "hit rate only" model each predict some but not all of the patterns? One possibility was that hit rate information was *more salient* than check rate information. The last four conditions shown in Figure 4 provided strong evidence for both high check rates (e.g., 6 bought out of 6 available) and low hit rates (e.g., 0 with a ticket out of 6 bought). When evidence for both was more or less equally strong, participants seemed to value the evidence of low hit rates over that of high check rates. In the first two conditions, there was strong evidence for low check rates (e.g., 1 bought out 6 available) but only weak evidence for high hit rates (e.g., 1 with a ticket out of 1 bought). It seemed only when hit rate information lacked strength did participants put emphasis on check rate information.

To test whether participants placed a higher weight on hit rate information, we fit a linear regression model using predictions generated by the "hit rate only" model and the "check rate only" model to predict human inferences. The estimated coefficient of the former (0.64) was greater than that of the latter (0.42), suggesting that participants may indeed have relied more on observed hit rates than check rates.

Of course, it's possible that some participants mostly relied on check rates and some on hit rates. In future research, we could use latent mixture models or similar techniques to identify subgroups and capture each subgroup's behavior.

General Discussion

Racial disparities have deep roots in societies around the world. Unfortunately, awareness doesn't always translate into the desire to end disparities. When reminded of racial disparities in prison, people became even more supportive of harsh criminal law that helped create them in the first place (e.g., Hetey & Eberhardt, 2014, 2018). Aside from stereotypic associations linking Blacks with crime, we argue that rational inference alone based on Naïve Utility Calculus (NUC) (Jara-Ettinger et al., 2016) can also reproduce disparities: If people believe police officers are knowledgeable about different races' "true crime rates" and act to maximize expected utilities (expected rewards of catching criminals relative to the total costs of checking), they may think groups under heavier scrutiny have higher crime rates and target them as well.

This is what we found in Experiment 1 when only check rates were available. To simulate police encounters without evoking stereotypes, we created a "Golden Ticket" game where robot chickens ("groups") lay eggs ("groups members") that may or may not have golden tickets ("crime") inside. Each ticket can be redeemed for a prize but each egg costs a token. Participants watched a knowledgeable, utility-maximizing agent Alex play a series of chickens, after which they were asked to infer each chicken's "hit rate" and decide whether or not to buy a new egg from it. While participants had no prior stereotypes linking certain chickens with high

or low hit rates, they reproduced disparities in Alex's sampling behavior nonetheless: The more eggs Alex bought, the more likely participants would buy from the same chicken. To show that participants were not copying without thinking, we asked them to infer each chicken's hit rate and their inferences closely matched a rational NUC model. Once again, these suggest that merely exposing people to disparities without addressing the underlying causes may well backfire. In Experiment 2, participants got to observe hit rates in the samples: Alex opened the eggs they bought in the end to reveal which ones had tickets and which ones didn't. Participants seemed to rely more on this new hit rate information than the agent's sampling behavior. Given strong evidence for high check rates (e.g., Alex bought 6 eggs out of 6) and low hit rates (e.g., 0 out of 6 eggs had tickets), participants expected true hit rates to be low. This was captured by a "hit rate only" that ignored check rates. When the evidence was strong for low check rates (e.g., bought 1 out of 6) but weak for high hit rates (e.g., 1 out of 1 egg had a ticket), participants inferred the true hit rates to be in a middling range. This inference was captured by the "check + hit rates" model.

A number of reasons may explain why participants seemed to value hit rate information more. Perhaps assuming that observed hit rates are equal to true hit rates takes less mental effort than decoding Alex's motivation (*"They wouldn't have bought so many eggs if this chicken weren't profitable."*). It could also be that participants doubted Alex's knowledge after a few "failures" (*"Maybe they haven't played with this one before?"*), even though they were reminded that Alex knew all chickens' hit rates really well. In follow-up studies, we can ask participants to rate Alex's credibility throughout the experiment to examine these hypotheses. Regardless of the underlying mechanism, findings in Experiment 2 suggested a way to fight disparities: Showing that groups targeted by the police are not necessarily more prone to crime may help people re-evaluate the true crime rates more objectively.

Future directions

Oftentimes, we don't just see one police officer repeatedly targeting certain groups but many across the nation targeting the same groups. To a naïve observer, consensus may be an even stronger indicator of "ground truth". In future work, we wish to look at whether hit rate information becomes less powerful when it goes against the consensus among multiple rational agents. Also, we used non-social groups in this study but it's likely that people reason about social groups differently. We will investigate this possibility in the future.

Acknowledgment

We thank Julian Jara-Ettinger, Tomer Ullman, Steve Piantadosi, Mahesh Srinivasan, the Berkeley Early Learning Lab, and four anonymous reviewers for their generous suggestions and thoughtful discussions that inspired and shaped this work.

Data and code used in this work are available at
`https://github.com/Yuan-Meng/PISG`

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *Fitting linear mixed-effects models using lme4.*

Del Toro, J., Lloyd, T., Buchanan, K. S., Robins, S. J., Bencharit, L. Z., Smiedt, M. G., ... Goff, P. A. (2019). The criminogenic and psychological effects of police stops on adolescent Black and Latino boys. *Proceedings of the National Academy of Sciences*, *116*(17), 8261–8268.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Hetey, R. C., & Eberhardt, J. L. (2014). Racial disparities in incarceration increase acceptance of punitive policies. *Psychological Science*, *25*(10), 1949–1954.

Hetey, R. C., & Eberhardt, J. L. (2018). The numbers don't speak for themselves: Racial disparities and the persistence of inequality in the criminal justice system. *Current Directions in Psychological Science*, *27*(3), 183–187.

Hetey, R. C., Monin, B., Maitreyi, A., & Eberhardt, J. L. (2016). *Data for change: A statistical analysis of police stops, searches, handcuffings, and arrests in Oakland, Calif., 2013-2014* (Tech. Rep.). Stanford University, SPARQ: Social Psychological Answers to Real-World Questions.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Sun, F., Schulz, L., & Tenenbaum, J. B. (2018). Sensitivity to the sampling process emerges from the principle of efficiency. *Cognitive Science*, *42*, 270–286.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–1140.

Liu, S., McCoy, J. P., & Ullman, T. D. (2019). People's perception of others' risk preferences. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 678–684). Montreal, QB: Cognitive Science Society.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). Cambridge: MIT Press.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668–676.