Naïve Utility Calculus underlies the reproduction of disparities in social groups

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

On the road to a more fair and just world, we must recognize ubiquitous disparities in our society, but awareness alone is not enough: Observed disparities between groups often get wrongly attributed to inherent traits (e.g., African Americans are disproportionately arrested because they are more prone to crime), creating a self-perpetuating feedback loop. As shown in a past study (Meng & Xu, 2020), such reasoning can result from the Naïve Utility Calculus (NUC, Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016): If an agent knows a target trait's "hit rate" in every group and avoids unnecessary sampling, it is rational to infer that groups sampled from more often have higher hit rates. The previous study used non-social categories (robot chickens) as stimuli, which raises the question of whether the results generalize to the social domain. In the current study, we replicated past findings using novel social groups (aliens): Overall, people were more likely to check groups examined more often by the agent but when observed hit rates did not support the agent's sampling behavior, people incorporated both information sources to infer group hit rates. This work brought NUC-based models one step closer towards tackling disparities in the real world consisted of social groups.

Keywords: social cognition; disparities; theory of mind; Bayesian models of cognition

Introduction

Quaerendo invenietis ("Seek and ye shall find").

— J. S. Bach, *Canon No. 9*

The arc of the moral universe may bend toward justice, but not by the awareness of injustice alone. Seeing, for instance, that Black men are 100 times more likely to get killed by the police than Asian women over the life course (Edwards, Lee, & Esposito, 2019), those in power may reinforce the *status quo*, possibly out of fear that groups disproportionately punished by the criminal justice system are indeed more prone to crime (Hetey & Eberhardt, 2014, 2018). Ironically, the more we scrutinize these groups, the more crime we discover, the more readily we justify our decisions, creating a feedback loop that forever amplifies accidental differences, or worse, deliberate biases (Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2017; Kearns & Roth, 2019).

So how can we break free from this vicious cycle? Hetey and Eberhardt (2018) suggested that we offer historical context for how disparities came to be, challenge stereotypes associating certain groups with certain traits, and highlight systemic forces behind disparities. While these measures are important and long overdue, they are not the end-all and be-all of the fight for social justice: Even without stereotypes and biases, disparities may still get perpetuated by low-level social cognitive processes. A recent study (Meng & Xu, 2020) examined how the "Naïve Utility Calculus" (NUC, Jara-Ettinger et al., 2016) alone can reproduce observed disparities: If we believe an agent (e.g., a police officer) knows the "hit rate" of a target trait (e.g., committing a crime) in each group and samples group members to check in a cost-efficient manner, we should infer that groups from which the agent sampled more often have higher hit rates; given the chance, we should also sample more from these groups. This is worrisome because we end up reproducing disparities in the agent's sampling behavior without prior stereotypes against any groups. The authors offered a solution, which was to show people what proportion out of group members checked by the agent actually had the target trait. They found that participants considered both sample hit rates and the agent's check rates when making inferences about groups: Groups that were checked often but had low sample hit rates or those rarely checked but had high sample hit rates were both thought to have moderate population hit rates. A key social implication of these findings is that, when exposing racial disparities in policing, we need to show both the number of police searches in different groups as well as the outcomes.

Social groups vs. non-social categories

While Meng and Xu (2020) attempted to examine disparities in society, the agent in their study sampled from robot chickens[1] rather than social groups. They used non-social categories as stimuli to prevent people from potentially bringing in existing stereotypes about actual social groups but this choice limits how well their findings generalize to social groups. It may well be the case that people reason differently about social groups and non-social categories.

One possibility is that social groups are considered more variable than non-social categories. For instance, even young children have strong intuitions that agents, not inanimate objects, can cause events to occur probabilistically—the latter

[1] In Meng and Xu's (2020) cover story, robot chickens lay eggs that may or may not contain a golden ticket and players can choose between buying an egg or letting it go. Participants watched a knowledgeable, utility-maximizing player interact with each chicken and were asked to infer the chicken's reward rate and decide whether to buy an egg from it. The game structure mirrored police encounters (player ↔ officer, chicken ↔ group, egg ↔ group member, reward ↔ crime).

tend to work as deterministic causes (Schulz & Sommerville, 2006; Wu, Muentener, & Schulz, 2016). In a similar vein, people may believe that the mental process behind crime is more subject to change than whatever mechanical process that controls reward rates in robot chickens. If a group of people are more unpredictable than a collection of robots, then sample hit rates may not generalize so well to populations.

However, one can also make a case for the opposite. *Social essentialism* (Rhodes & Moty, 2020) may lead us to think members of a meaningful social group share an "essence", which cannot be said about eggs that merely share the space inside the same robot chicken. In this case, sample hit rates in essentialized social groups should generalize better to populations than those in non-essentialized non-social categories.

A final possibility is that, despite potential differences between social groups and non-social categories, the same inferential process underlies how we infer population hit rates from an agent's sampling behavior and sample hit rates.

Goals of current research

The primary goal of the current research is to examine whether findings in Meng and Xu (2020) generalize to social groups. If we find no major domain differences, it brings NUC-based models one step closer towards explaining and tackling real-world disparities. If the NUC does not capture people's hit rate inferences in social groups, then future studies are needed to find formal accounts in the social domain.

Apart from the nature of the groups, several aspects in Meng and Xu (2020) can also be made more true to life. For instance, in the robot chicken game, the trait that the agent was looking for was positive (i.e., containing a reward) whereas in police encounters, the target trait is most definitely negative (e.g., committing a crime). A plethora of research suggests that people tend to weigh negative outcomes more heavily than positive ones (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Fazio, Eiser, & Shook, 2004). So if letting a criminal roam free is more consequential than winning a reward in a game, an agent should adopt a lower threshold for when to check in the former case—as a result, the same observed check rate would imply a lower population hit rate. To match outcome valence in police encounters and similarly high-stake scenarios, we used negative outcomes in the current study. Another major concern about Meng and Xu's (2020) study is that there were as as many "critical trials" in which sample hit rates contradicted the agent's check rates as there were "control trials" in which the two rates in agreement. Upon seeing so much counterevidence, people may no longer trust the agent, which breaks the core assumptions of the NUC about the agent's knowledge and efficiency, making it tricky, if not impossible, to infer population hit rates from the agent's sampling behavior. To protect the agent's "reputation', our study included more control than critical trials.

Overview of experiment

To achieve the aforementioned goals, we created a "Gem Patrol" game, which was structurally similar to the "Golden Ticket" game in Meng and Xu's (2020) study. Our game is set on the planet Obos that produces a rare and precious gem. Aliens from nearby planets often visit Obos to buy gems, some of whom attempt to leave without paying. Unpaid gems emit a signal that can be detected by a special machine, which runs on expensive reagents and can be slow, making scanning costly. Border patrol officers decide whether to scan each visitor or let them go directly. Alex is said to be the best officer because they know the "theft rate" of aliens from each planet and only scan visitors if necessary. Participants watched Alex interact with visitors from a series of planets. After each planet, they were asked to infer the theft rate of all aliens on that planet and decide whether to scan a new visitor from there. Figure 1 illustrates an example trial. In the "Check Rate Only" condition, scan results were never shown to participants so they could only infer theft rates based on Alex's check rates. In the "Check + Hit Rates" condition, participants got to see how many visitors actually stole a gem.

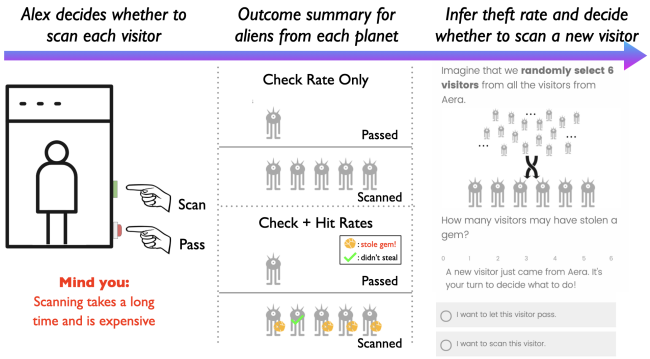

Figure 1: The "Gem Patrol" game: In each trial, a group of 6 aliens from the same planet (e.g., Aera) went through border security before leaving planet Obos. Some may attempt to leave with stolen gems, which can be detected by a special machine. Border patrol officer Alex knows the theft rate of aliens from each planet and only scans visitors if necessary. In the "Check Rate Only" condition, participants saw how many visitors Alex scanned but not the results. In the "Check + Hit Rates" condition, participants also saw how many stole a gem out of those scanned. At the end of each trial, participants were asked to infer the theft rate of all aliens on a planet and decide whether to scan a new visitor from there.

Computational Modeling

Given data in a sample, how should people infer the theft rate, or more generally, the hit rate of the population? To answer this question, we adopted computational models in Meng and Xu (2020), which are shown graphically in Figure 2.

NUC-based models

One class of models are based on the NUC (Jara-Ettinger et al., 2016), which assumes the agent Alex maximizes expected utilities when deciding whether to scan a visitor. This assumption creates a link between Alex's check rate μ of aliens

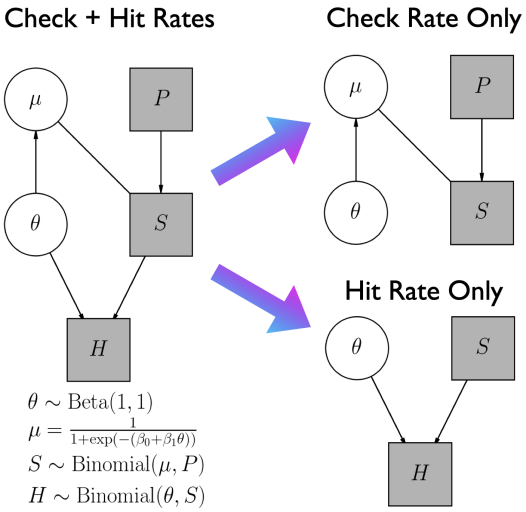

Figure 2: Computational models for inferring population hit rates. In Naïve Utility Calculus (NUC) models, the agent Alex's check rate μ for aliens from a given planet is linked to their population hit rate θ via a logistic function, whose intercept β_0 and slope β_1 are estimated from population hit rates inferred by participants and their scanning decisions. μ is inferred from the number of visitors Alex scanned (S) out of the total number of visitors they encountered (P). In the "Check Rate Only" condition, θ can only be inferred from μ ("Check Rate Only" model). In the "Check + Hit Rates" condition, θ can also be simultaneously inferred from μ and the number of gem thieves (H) out of P ("Check + Hit Rates" model). The non-NUC "Hit Rate Only" model ignores μ and infers population hit rates solely from sample hit rates.

from a given planet and the population hit rate θ on that planet. If each scan costs c and catching a gem thief returns a reward r, then the expected utility of scanning a visitor is:

$$\mathrm{E}[U(scan)] = r\theta - c. \tag{1}$$

$\mathrm{E}[U(scan)]$ is then connected to μ via a logistic function:

$$\mu = \frac{1}{1+e^{-\mathrm{E}[U(scan)]/\tau}}, \tag{2}$$

where the temperature parameter τ captures Alex's decision noise: Under extremely low temperatures ($\tau \to 0$), Alex will scan a visitor as long as $\mathrm{E}[U(scan)] > 0$; when temperatures are extremely high ($\tau \to \infty$), Alex randomly chooses between whether or not to scan. The logistic function is a special case of the softmax function (under binary choices), which is commonly used to convert unbounded action values into action probabilities bound by 0 and 1 (Sutton & Barto, 1998).

By plugging Equation 1 into Equation 2, we can rewrite the original logistic function as follows:

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

where the slope β_1 is r/τ and the intercept β_0 is $-c/\tau$. Both parameters can be empirically estimated from participants' hit rate inferences ("Out of N random new visitors from this planet, how many may be thieves?") and their scanning decisions ("Do you want to scan a visitor from there?"). Estimates of β_0 and β_1 can be plugged back in Equation 3 to infer θ from μ; μ itself is inferred from how many visitors Alex scanned, S, out of P visitors from a given planet—$S \sim \text{Binomial}(\mu, P)$.

In the "Check Rate Only" condition, the "Check Rate Only" model is the only one for population hit rate inference. In the "Check + Hit Rates" condition, learners can also take account of sample hit rate information, simultaneously inferring θ from μ as well as the number of gem thieves, H, out of S visitors that Alex scanned—$H \sim \text{Binomial}(\theta, S)$. Such an inferential process is captured by the "Check + Hit Rates" model. Alternatively, learners may ignore hit rate information and still rely on the "Check Rate Only" model like before.

Non-NUC model

Both the "Check Rate Only" and the "Check + Hit Rates" models hinge on the agent's utility maximization. However, learners may turn a blind eye to the agent's sampling behavior and infer population hit rates solely based on sample hit rates. This type of learning is captured by the non-NUC "Hit Rate Only" model, which infers θ from the number of gem thieves, H, out of S visitors Alex scanned—$H \sim \text{Binomial}(\theta, S)$.

Experiment

Methods

Participants We recruited 145 participants living in the United States from Amazon Mechanical Turk. Each participant gave informed consent before participating and was paid \$2.5 for about 15-20 minutes of their time. Sixty-three[2] were excluded from analysis for failing to answer all comprehension check questions correctly. Among the remaining 82, 46 (mean age = 34.28 years, SD = 9.5) were in the "Check Rate Only" condition and 36 (mean age = 34.31 years, SD = 8.2) were in the "Check + Hit Rates" condition.

Design Table 1 summarizes the trial content. In the "Check + Hit Rates" condition, there were 6 critical trials where sample hit rates contradicted Alex's check rates and 10 control trials where sample hit rates and check rates were agreed with one another (both high, low, or moderate). In all trials, the total number of visitors Alex encountered was fixed to be 6. In the "Check Rate Only" condition, participants only saw how many visitors Alex scanned in each trial but not how many of them stole a gem. Participants were randomly assigned to one of the two conditions and the trial order was randomized for each person.

[2]The high attrition rate (43.4%) may be a result of the COVID-19 pandemic: After the imposition of sheltering in place and an influx of new workers, average workers in the Amazon Mechanical Turk pool were less attentive than before (Arechar & Rand, 2020). To address the concern that the sample here was biased after exclusion, we replicated this experiment as part of a follow-up study on Prolific and found similar results at a much lower attrition rate (12.0%).

Table 1: Summary of trial content. The total number of visitors Alex encountered was always 6 in all trials. The number of gem thieves was only revealed in the "Check + Hit Rates" condition but not in the "Check Rate Only" condition.

trial type	# scanned	# thieves
critical	6	1
	6	0
	5	1
	5	0
	2	2
	1	1
control	6	6
	6	5
	5	5
	5	4
	4	3
	4	2
	3	2
	3	1
	2	0
	1	0

Procedure To begin, participants watched a short video introducing the "Gem Patrol" game as previously described and were tested on the utility structure of the game (costs, rewards), what makes Alex the best border patrol officer (that they know the theft rate of aliens from each planet and maximize expected utilities of scanning), and randomness in small samples (that 6 visitors may not represent aliens on an entire planet). Only those who answered all the comprehension check questions correctly were allowed to continue.

Each trial began with a picture of 6 aliens from one planet lining up to pass border security. Those from each planet had unique body colors. The next page depicted how many visitors Alex scanned or let pass, which was the only information available in the "Check Rate Only" condition. In the "Check + Hit Rates" condition, it was also revealed how many stole a gem out of those scanned. In both conditions, participants answered two questions at the end of each trial. The first was that, out of 6 random new visitors from a given planet, how many might be gem thieves. This measured participant's inferences of population hit rates. The second question was whether they wanted to scan a new visitor from that planet, the answers to which were used to estimate free parameters β_0 and β_1 in NUC-based models.

Results

Did people copy Alex's sampling behavior? As shown in Figure 3 (left), overall, people followed Alex's footsteps: In both conditions across all 16 trials, the higher the proportion of visitors Alex scanned out of the total from a given planet, the more likely participants would choose to scan a new visitor from there. To test this observation, we fit two general-

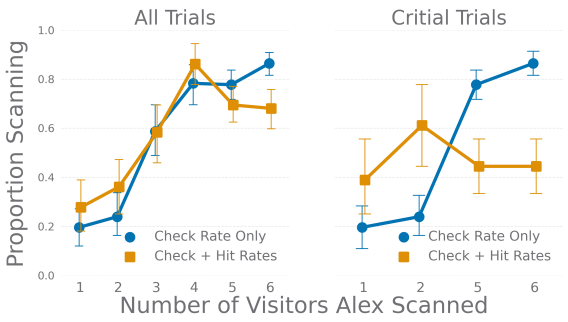

Figure 3: Reproduction of disparities: The proportion of participants deciding to scan a new visitor as a function of how many visitors Alex scanned in all trials (left) and critical trials (right). (Error bars indicate the 95% confidence intervals.)

Figure 4: Population hit rates inferred by people vs. models: In the "Check Rate Only" condition (upper), the "Check Rate Only" model is the only hit rate inference model. In the "Check + Hit Rates" condition (lower), the "Check + Hit Rates" model takes account of both check and hit rates whereas the other two models rely on one of the two. (Solid black bars indicate the 95% confidence intervals in human inferences. Dashed gray bars indicate the 95% highest density intervals of posterior distributions in Bayesian models.)

ized linear mixed models (GLMMs) with data from each of the two conditions using the Python package `Pymer4` (Jolly, 2018): Participants' scanning decisions (1 = scan; 0 = let pass) were the outcome variable and the number of visitors Alex scanned (out of 6) was the fixed effect; this model has random intercepts and slopes for participants and trials[3]. As expected, the fixed effect was significant in the "Check Rate Only" condition: With each additional visitor Alex scanned, the odds ratio for participants scanning a visitor increased by 2.19, which was significantly higher than chance, Wald's χ^2 = 14.98, $p = .00011$. This was also true in "Check + Hit Rates" condition where sample hit rates contradicted Alex's check rates in several trials: With each additional visitor Alex scanned, the odds ratio for participants scanning a visitor increased by 1.60, which was significantly higher than chance, Wald's χ^2 = 5.12, $p = .023$. These results showed that, in general, participants copied Alex's sampling behavior, even in the face of occasional contradictory information.

However, on the 6 critical trials where sample hit rates were inconsistent with Alex's check rates, did participants still copy Alex blindly? According to Figure 3 (right), this did not seem like the case. To test this new observation, we fit another GLMM with critical-trial data in the "Check + Hit Rates" condition. The fixed effect was no longer significant: With each additional visitor Alex scanned, the odds ratio for participants scanning a visitor decreased by .03, which was not distinguishable from chance, Wald's χ^2 = .17, $p = .68$.

How did people infer population hit rates? As with Meng and Xu (2020), a key assumption behind this work is that people's scanning decisions are driven by inferred population hit rates. Figure 2 illustrated three different inferential processes, each of which was implemented by `PyMC3` (Salvatier, Wiecki, & Fonnesbeck, 2016), a Bayesian modeling library in Python. For NUC-based models ("Check Rate Only" and "Check + Hit Rates"), we estimated[4] parameter values of β_0 and β_1 in Equation 2 using people's inferred hit rates and scanning decisions. The estimates were $\beta_0 = -1.98$ and $\beta_1 = 4.67$ in the "Check Rate Only" condition and $\beta_0 = -2.87$ and $\beta_1 = 9.48$ in the "Check + Hit Rates" condition, respectively. The non-NUC model ("Hit Rate Only") has no free parameters.

In all three models, population hit rates are derived from first principles rather than inferred from human data, so model comparison metrics such as log-likelihood, AIC, BIC, etc. do not apply here. To compare our models, we can examine how well each model's predictions align with human inferences by looking at Pearson's r and root-mean-square error (RMSE). As shown in Figure 4 (upper), in the "Check Rate Only" condition, population hit rates inferred by people were qualitatively similar to predictions of the "Check Rate Only" model. The two were strongly correlated, Pearson's $r = .55$, $p < .001$. In the "Check + Hit Rates" condition, all model inferences correlated with human inferences: The "Check + Hit Rates" model the most strongly, Pearson's $r = .66$, the "Hit Rate Only" model the least so, Pearson's $r = .41$, and the "Check Rate Only" model in between, Pearson's $r = .58$. We can also use RMSE to measure how much each model's predictions deviated from human inferences; lower values suggest less deviation. In this regard, the "Check + Hit Rates" model was the closest to humans among the three, RMSE = .23, the "Hit Rate Only" model the furthest, RMSE = .29, and the "Check Rate Only" model once again in between, RMSE = .26. According to both metrics, the "Check + Hit Rates" model best captured human inferences, as shown in Figure 4 (lower). Upon a closer look, the "Check Rate Only" model overestimated population hit rates when Alex scanned a large proportion of visitors, even though sample hit rates were very low. The "Hit Rate Only" model made the opposite mistakes, overestimating population hit rates when sample hit rates were high, even though Alex scanned few visitors. By comparison, like people, the "Check + Hit Rates" model inferred moderate population hit rates to reconcile conflicting check rates and hit rates in the samples.

Discussion

The current study replicated Meng and Xu's (2020) findings in the social domain. First of all, if a knowledgeable, utility-maximizing agent frequently sampled members from a social group to check at a cost, participants also tended to check a new member from this group. This pattern was found in both the "Check Rate Only" and the "Check + Hit Rates" conditions across all trials, suggesting that people generally trusted the agent and reproduced their sampling behavior, despite occasionally receiving contradictory information (i.e., sample hit rates were low in groups that the agent checked often but high in groups that the agent did not check much). Importantly, however, on critical trials where sample hit rates deviated from the agent's check rates, people no longer copied the agent's behavior. This finding suggested that providing information on sample hit rates may be an effective way to curb the reproduction of disparities in the social domain.

We argue that people's sampling decisions are informed by their inferences about population hit rates in different social groups. To examine the underlying inferential process, we implemented three Bayesian models learning from different sources of information ("Check Rate Only", "Hit Rate Only", and "Check + Hit Rates") and compared their predictions with people's hit rate inferences. Among these, the "Check + Hit Rates" model resembled humans the most and captured major patterns in our participants' inferences, suggesting that people may combine both sample hit rates and the agent's check rates when inferring population hit rates.

General Discussion

Disparities between different demographical groups are deeply rooted in all aspects of our society, from education,

[3]The formula of all GLMMs used in this paper is `scan ~ n_checked + (1 | participant) + (1 | trial)`.

[4]Meng and Xu (2020) obtained *maximum a posteriori* (MAP) estimates of β_0 and β_1 before feeding them to NUC models. Here the choice model is built into NUC models so posterior distributions of parameter values, not MAP estimates, are used to infer hit rates.

employment to public health, criminal justice, just to name a few. How we interpret evidence of disparities is like the Rorschach test: Some attribute observed disparities to individual or systemic biases and some to traits inherent to the groups. The latter impedes progress if people believe certain groups are to blame when historical data are distorted by self-fulfilling prophecies. A recent study (Meng & Xu, 2020) provided a formal account to explain why we may attribute disparities to inherent group traits and offered a potential solution. The explanation was based on the Naïve Utility Calculus (NUC, Jara-Ettinger et al., 2016): If an agent knows the hit rate of a target trait in different groups and avoids sampling (e.g., hiring, arresting, paroling) group members unnecessarily, then it is rational to infer that groups sampled from more often are more likely to possess that trait. The solution was to show people the hit rates in the samples so they could combine this new information with the agent's sampling behavior to adjust inferences about true trait prevalence in groups.

Critically, however, the previous study stripped away the social context of groups that the agent sampled from, using robot chickens so that participants could not rely on stereotypes they might have about actual groups. To see whether past results could generalize to the social domain, we designed a new study in which the agent sampled from novel social groups (e.g., aliens from different planets), thereby preserving the social nature of groups while keeping away existing stereotypes. We replicated Meng and Xu's (2020) findings that, overall, people tended to copy a knowledgeable, utility-maximizing agent's sampling behavior but not when sample hit rates contradicted the agent's check rates. In the latter case, participants relied on both the sample hit rates and the agent's check rates to infer groups' hit rates and decided whether to check a new group member accordingly. The current results brought NUC-based explanations and solutions closer towards the real world consisted of social groups.

Future directions

From alien planets to the world we live in, there is a long way to go. Ultimately, we wish to understand why it is that awareness of disparities can sometimes lead to future disparities and, most importantly, find ways to stop this vicious cycle of injustice. There are many directions in which we can go.

First of all, the current NUC models operate on the premise that the agent is always knowledgeable, cost-efficient, and well-intentioned, which often cannot be further from the truth, such as when former police officer Derek Chauvin suffocated George Floyd to death, whose only crime was using a 20-dollar counterfeit bill at a grocery store. Challenging people's blind trust in the agent is potentially another effective way to stop the reproduction of disparities. To this end, we need models that jointly update how learners think about groups and how much they still trust the agent after receiving evidence contradicting the agent's knowledge and efficiency.

Another way to make NUC models more realistic is to incorporate uncertainty into the choice model. That is, if some groups are rarely sampled from, we should know little about their true hit rates; as such, the expected utility of checking is highly uncertain and we should probably discount this information when making critical decisions (e.g., Li & Ma, 2020).

Moreover, we plan on using actual social groups in follow-up studies. The merit of using novel social groups is that we can steer away from people's existing stereotypes; however, it is real social groups that we care about—How will people think and act when new statistical information clashes with their prior social knowledge? Do they still rely on the same NUC to draw or update inferences about groups? Or, does the answer depend on what kind of prior knowledge they have and other social factors (e.g., political positions, demographics)? These are important questions to investigate next.

Last but not least, we hope to examine the developmental origin of the current results: Do children also reproduce disparities that they see without prior biases? Does the NUC also underlie their decisions and inferences about social groups?

Acknowledgment

We thank Julian Jara-Ettinger, Steve Piantadosi, Mahesh Srinivasan, Tomer Ullman, the Berkeley Early Learning Lab, and two anonymous reviewers for their helpful suggestions, discussions, and feedback. We also thank the audience members at SRCD 2021 for many thought-provoking questions when we presented a modified version of this work there.

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

References

Arechar, A. A., & Rand, D. (2020, November 27). Turking in the time of COVID. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/vktqu`

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Edwards, F., Lee, H., & Esposito, M. (2019). Risk of being killed by police use of force in the united states by age, race–ethnicity, and sex. *Proceedings of the National Academy of Sciences*, *116*(34), 16793–16798.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. E., & Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. *CoRR*, *abs/1706.09847*. Retrieved from `http://arxiv.org/abs/1706.09847`

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293-311.

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.

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.

Jolly, E. (2018). Pymer4: Connecting R and Python for linear mixed modeling. *Journal of Open Source Software*, *3*(31), 862.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford, England: Oxford University Press.

Li, Z., & Ma, W. J. (2020, October 8). An uncertainty-based model of the effects of fixation on choice. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/ajmwx`

Meng, Y., & Xu, F. (2020). How do disparities reproduce themselves? "Ground truth" inference from utility-maximizing agent's sampling behavior. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 903–908). Austin, TX: Cognitive Science Society.

Rhodes, M., & Moty, K. (2020). What is social essentialism and how does it develop? In M. Rhodes (Ed.), *The development of social essentialism* (p. 1-30). Cambridge, MA: Academic Press.

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

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, *77*(2), 427–442.

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

Wu, Y., Muentener, P., & Schulz, L. E. (2016). The invisible hand: Toddlers connect probabilistic events with agentive causes. *Cognitive Science*, *40*(8), 1854–1876.