Fair play in research and policy

A paper in this issue identifies a persistent influence of irrelevant information in social contexts, which results in biased and unfair judgements. These widespread social biases can be insidious as they inadvertently enter research and policy.

The past few decades have seen a proliferation of documented human biases in psychology research; from cognitive to social domains, the tendency to act or think in ways that systematically deviate from normative or rational standards of judgement appears to be pervasive. Judgements based on perceived statistical regularities may simplify a complex world, facilitating behaviour, but can also lead to inaccuracy in individual instances. In social domains, these tendencies take on a more sinister valence as they can result in prejudice against certain groups and pit perceived statistical regularities against considerations of fairness. And their effects can be ubiquitous, even permeating quantitative data and other measures thought to carry objective weight. Awareness of these tensions is particularly important at the level of policy, where social biases can have widespread impacts and be further compounded by the fact that select groups have asymmetric power to shape law, policy and research agendas.

In this issue, Jack Cao and colleagues (article no. 0218) highlight the pervasiveness and stubbornness of a social bias they call base rate intrusion. Base rates are the prior probability of a group having a particular trait (for example, rates of breast cancer in women versus men), which can be important information for making predictions but can also engender attitudes and behaviour that vary with group membership¹. This sort of problem pits perceived likelihood against what is considered to be morally fair (or, in some cases, legal). The authors focus on a specific type of base rate — stereotypes about social groups — and construct a series of social (and, for comparison, non-social) scenarios for which base rate information is inconsequential. For example, their scenarios involve doctors and nurses, who tend to be male and female, respectively, and, for comparison, broken and intact spoons, which tend to be plastic or metal, respectively. They show that base rates have a pernicious influence, resulting in inaccurate and unfair judgements even when

Credit: Arturbo / E+ / Getty Images

irrelevant, but only in social domains (for example, involving race, gender, nationality).

In general, disparate outcomes between social groups are common. However, obtaining objective measures of these differences is complicated by social biases and incorporating them into decisions can have significant ramifications for marginalized groups. Despite their potential predictive capacity, the use of base rates can violate tenets of fairness and justice; for this reason, their use in US courts has been contested. For example, take the use of recidivism 'risk scores' computed by an algorithm to inform bail decisions (though the final decision is ultimately up to the judge)^{2,3}. Implicit in their use is the belief that data represent fact and that algorithms are less susceptible to social biases than are judges.

But data can reflect the social biases of those who generate or collect them. For example, arrest outcomes depend on the discretion of patrol officers, who decide which individuals to investigate and whether a behaviour constitutes a crime. In the United States, for instance, even though black and white people use cannabis at similar rates, black people are almost four times more likely to be arrested for cannabis possession than whites⁴. And algorithms are only as objective as the data they are built on. A recent study showed that algorithms trained on large corpora of human text inherit many of our implicit social stereotypes⁵, which would have deleterious consequences if they were given agency over decisions. In addition to the concern that

the data informing base rates can itself be biased is the potential that their use results in cyclical patterns, whereby differences in base rates are perpetuated in vicious cycles. For example, the proactive targeting of certain high-crime neighbourhoods by police might result in inflated differences in rates of arrest between groups, which may in turn perpetuate the conditions that lead to increased crime in these neighbourhoods.

Cao et al. show that in social contexts, relying on base rates can both harm decisions and be unfair. They also show that the tendency to rely on stereotypes is resistant to change and only corrected when all references to base rates have been removed and persons individuated. Despite the stickiness of the tendency to stereotype, that base rate intrusion can be ameliorated at all provides a glimmer of hope and a potential path to intervention.

As long as there are disparate outcomes between social groups, the question of how to account for these differences will have to be sensitive to our own tendencies as well as to the disproportionate impacts they have on certain individuals when they enter legal and political frameworks. This requires acknowledging that data does not represent objective truth and that notions of what is fair are central. Understanding the nature and underlying causes of social biases, their insidious and widespread effects, and how group differences can be recognized fairly is an important pursuit with vast implications for research and policy. □

Published online: 10 October 2017
DOI: 10.1038/s41562-017-0231-1

References

1. Kahneman, D. & Tversky, A. *Psychol. Rev.* **80**, 237–251 (1973).
2. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (23 May 2016).
3. Lum, K. *Nat. Hum. Behav.* **1**, 0141 (2017).
4. Edwards, E., Bunting, W. & Garcia, L. *The War on Marijuana in Black and White* (American Civil Liberties Union, New York, 2013).
5. Caliskan, A., Bryson, J. J. & Narayanan, A. *Science* **356**, 183–186 (2017).