6

SOCIETY MADE ME DO IT

INTRODUCTION

The representativeness heuristic comes into play when we rely on stereotypes to make judgements such as the probability that a particular person from a defined group has a particular attribute. A stereotype is a collection of associated attributes learned from experience and through cultural transmission. 'Having a stereotype doesn't make you a racist, sexist, or whatever-ist. It just means your brain is working properly, noticing patterns, and making generalizations' (Payne, Niemi, & Doris, 2018). It is, therefore, very useful but sometimes misleading. Thus, if all the plumbers you have encountered are male, then maleness would be a default attribute when you think of a plumber. Much of the time, this would be a useful culturally derived heuristic which is also indistinguishable from a bias (Hinton, 2017a, 2017b). Representativeness can work in two directions. An individual from a particular group is deemed to inherit the attributes of the group stereotype and hence represent it (see [a] in Figure 6.1). Alternatively, the attributes of a person from a group one is not particularly familiar with can be assumed to represent the attributes of the group (see [b] in Figure 6.1).

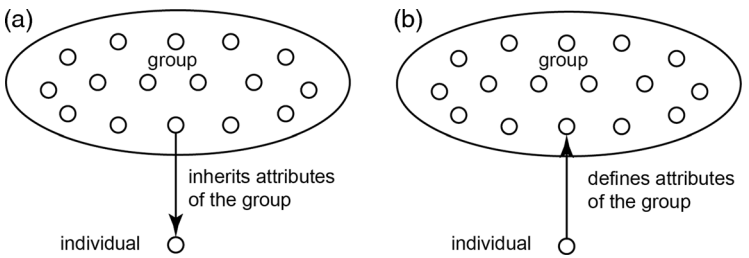

FIGURE 6.1 In (a), an individual from a particular circumscribed group is deemed to inherit the presumed attributes of the group (a stereotype). In (b), the attributes of an individual are deemed to be representative of an unfamiliar group the individual comes from.

For example, if you meet a grumpy Bulgarian with no sense of humour, you might conclude: 'Gosh, aren't Bulgarians grumpy and humourless'. This is an example of the 'liberal induction' we encountered in Chapter 4. Similarly, if a member of a group does something blameworthy, then the group in general may be seen as collectively to blame as often happens, for example, to Muslims. Chinese students in the UK and mainland Europe received racist comments or were even attacked because of the coronavirus pandemic. They somehow 'inherited the attribute' of having the virus due to being part of the same ethnic group as those Chinese citizens 5,000 miles away who had the virus. In April 2020, ISIS gunmen and suicide bombers killed 25 Sikhs in a temple in Kabul in Afghanistan in revenge for India's treatment of Muslims in Kashmir. The Sikhs in Afghanistan inherited the collective blame for what was going on in India. Bias due to stereotyping can affect an individual's judgement of others based on their ethnicity, gender, accent, class, religion, university education, and so on; and judgements of others can have an impact on the success or failure at job interviews or whether they are likely to be stopped and searched by the police. Such biases are often deemed implicit or unconscious and, to ensure that inadvertent discrimination does not take place, there is a perceived need to

address them through some form of intervention. To support this endeavour, there are many websites and company development sessions offering 'diversity training' aimed at helping people overcome their unconscious biases, with race being the primary attribute that tests tend to measure.

There are some assumptions about cognitive biases, both in some studies by Tversky and Kahneman and among those engaged in diversity training. These include the assumption that:

- They can be measured;
- Biases are unconscious and that we respond intuitively to information that comes most immediately and automatically to mind;
- We all possess biases due to the ways our cognitive system operates;
- They form the basis of much of our attitudes and behaviour;
- Through diversity training, implicit racial bias can be reduced.

A first stage in some types of diversity training is to measure the strength of any unconscious bias a person might have. This is generally done using an implicit association test (IAT) of some kind (Greenwald, McGhee, & Schwartz, 1998) that measures latencies (slight delays) in responding to words or pictures presented on a computer. Participants might be presented with words for flowers such as 'rose' or musical instruments such as 'violin' on a computer screen and asked to respond as quickly as possible using the word 'pleasant', and to respond 'unpleasant' to insect words such as 'wasp' or weapons such as 'hatchet'. One might predict that such responses should be quite easy given the nature of the nouns and the attributes implicitly attached to them – such noun-attribute combinations are deemed compatible. There are various ways in which the instructions for this type of measure can be modified; for example, participants might be asked to reverse the associated noun-attribute so that they have to respond 'unpleasant' to 'daffodil' or 'pleasant' to 'gun'. In this case, because these

combinations are assumed to be incompatible, the time taken to respond should increase compared to the compatible condition. Results from experiments by Greenwald et al. (1998) showed that significant differences in the latencies between compatible and incompatible word conditions provides evidence that we implicitly associate certain categories of things as being pleasant or unpleasant (experiment 1).

In Experiment 2, they extended this method to assessing evaluative associations between Japanese Americans and Korean Americans since each group were deemed to have negative attitudes towards the other (out-group) due to the historical occupation of Korea by the Japanese in the first half of the 20th century. The stimuli in this case were typical Japanese names and typical Korean names. As expected, the Korean American participants responded more quickly with 'pleasant' to Korean names than to Japanese names, and with a longer latency when responding 'pleasant' to Japanese names.

A third experiment looked at the patterns of attitude difference between Black (African American) and White (European American) subjects again using names typically associated with each group. In this experiment, there was a strong association between White names and 'pleasant' rather than Black names and 'pleasant'. Furthermore, the study included other measures used to compare and evaluate the IAT, the Modern Racism Scale (McConahay, Hardee, & Batts, 1981), and Diversity and Discrimination scales (Wittenbrink, Judd, & Park, 1997). Based on the answers given to these scales, the majority of the White participants had expressed no negative feelings towards African Americans, essentially a Black-White indifference, nevertheless their scores on the IAT suggested a White preference. Thus, the IAT was, arguably, tapping into unconscious bias.

WHERE DO THE BIASES COME FROM?

The biases assessed in diversity training come about because our judgements are based on the ways we have learned to categorize objects, events, and people in the past. While some stereotypes might come from personal experience, prejudice

such as racism is rarely acquired that way but rather through what people learn from their sociocultural context in childhood, from their peers, and from the media. A number of cognitive social psychologists have suggested that prejudice was an inevitable outcome of such categorization processes (Beck & Forstmeier, 2007; Devine, 1989; Tajfel, 1981) and, therefore, that dealing with prejudice is likely to be difficult. As cited in Billig (1985), Tajfel (1981, p. 141) asserted that 'there is no easy way to deal with intergroup prejudice in its manifold varieties, and all one can hope for is that its more vicious and inhuman forms can be made less acute sooner or later'. Billig regarded this view as a form of fatalism, and that the categorization of a particular stimulus could be countered by *particularization* where a particular stimulus, an individual in this case, is treated as a special case. In his view, tolerance of others through this particularization process has not been sufficiently considered by the categorization psychologists. However, the argument that one can deal with implicit bias by focussing on particular cases seems to be a version of the 'no true Scotsman' fallacy where someone, who does not seem to meet all the putative criteria of a category, is considered to be not a true member of the category. The argument goes: 'No Scotsman puts sugar in his porridge.' 'But Hamish puts sugar in his porridge.' 'No *true* Scotsman puts sugar in his porridge'. Thereby, the particular case of Hamish is dismissed from the category of true Scotsmen and thus the category itself (the stereotype) lives on.

The idea that prejudice is an intractable problem continues to this day, and so, when it comes to racial, gender, educational bias, etc., there are those who argue that diversity training is ineffective. Bezrukova, Spell, Perry, & Jehn (2016) concluded that there was 'no compelling evidence that long-term effects of diversity training are sustainable in relation to attitudinal/affective outcomes' (p. 1243). Noon (2017) questions the usefulness of applying the results of psychological tests of bias to diversity training and also argues that they are generally ineffective. He argues that it is possible that 'aversive racists' only (those who are sure they are not at all prejudiced but who have

suppressed negative beliefs about racial minorities) are likely to be influenced by unconscious bias training. Measuring bias through IATs can have two potentially contradictory effects: first, to break the prejudice habit 'people must be aware of their biases and, second, they must be concerned about the consequences of their biases before they will be motivated to exert effort to eliminate them' (Devine, Forscher, Austin, and Cox 2012, p. 1268). Second, tests of bias that reveal underlying unconscious prejudice provide a way of avoiding blame since, being deeply embedded and a result of normal human categorization processes, the bias is 'not our fault'.

OVERCOMING BIASES

So, given the results of IATs, what aspects of the human cognitive system need to be addressed? Devine (1989) has referred to stereotypes as a form of habit derived from well-learned associations that are, in turn, caused by 'repeated activation in memory'. This will vary from person to person resulting in low to high levels of prejudice. Stereotyping, and hence implicit bias, is, therefore, a form of automatic and largely unconscious (heuristic) processing. According to her, when one's personal beliefs overlap with a negative stereotype, then there is the potential for explicit prejudice. Where one's personal beliefs do not overlap with the stereotype, then there is a degree of conflict since cultural stereotypes are automatically triggered for both low and high prejudiced individuals. Overcoming this habit requires conscious controlled processing. A prejudice triggered by the activation of a stereotype is the result of Type 1 thinking and is enduring and difficult to counter. On the other hand, controlled processes can cause the explicit system (Type 2 thinking) to change relatively quickly – it is 'malleable'. But, despite short-term apparent changes to expressions of explicit bias, they are likely to be short-lived.

Are there interventions that work to produce long-term changes to implicit bias; or at the very least that can mitigate the expression of bias? One way is to compare a variety of

interventions to find out which ones work. Another is a scattergun approach where a range of strategies is used in a single intervention. As an example of the former strategy, Lai et al. (2014) describe an unusual research contest where a number of teams were asked to come up with various approaches that could be tested against others to see which ones, if any, reduced implicit racial preferences. The incentive for the teams of researchers who took part was to win the contest. A total of 18 interventions plus a control were used coming under the overall headings of: 'Exposure to counter-stereotypical exemplars', 'Intentional strategies to overcome biases', 'Evaluative conditioning' (in this case, pairing words: Black with positive attributes and Whites with negative attributes), 'Appeals to egalitarian values', 'Engaging with others' perspectives', and 'Inducing emotion'. Of these, the first three were effective at reducing implicit preferences and the last three ineffective. While the contest did not examine how long lasting the strategies were likely to be, the conclusion was that interventions that use multiple strategies are likely to be effective. In a follow up study, Lai et al. (2016) looked at the degree to which implicit racial preferences were reduced over time following the interventions that had shown promise. The time period over which they were tested was several hours to several days. Although the interventions showed that implicit preferences were malleable in the short term, none led to long-term change.

In contrast, an earlier study by Devine et al. (2012) claimed to produce a sustained reduction in bias over a three-month period. Their method was not to compare which method was most effective but rather to provide a variety of bias reduction strategies (five in all) in the hope that altogether they would induce a degree of self-regulation on the part of the participants. Their aim was to translate their 'situational awareness into chronic awareness of biases in themselves and in society, thereby flipping the self-regulatory switch that motivates strategy use and reduces implicit bias' (p. 1268). Based on their results, they make the case that offering a variety of strategies allows for individual differences in what is likely to motivate an individual to reduce expressions of bias by stimulating Type 2

thinking in situations that are likely to activate the (Type 1) 'habit' of bias. 'Our data provide evidence demonstrating the power of the conscious mind to intentionally deploy strategies to overcome implicit bias' (Devine et al. 2012, p. 1277).

Studies by Bruneau, Kteily, & Urbiola (2019) provide further examples of combating racial or religious bias by tapping into *collective blame hypocrisy*. This occurs when people blame an out-group (Muslims in this case) in its entirety for the bad actions of individual group members, but do not blame their own group in its entirety for bad actions by in-group members. Their method taps into our desire for consistency in our thoughts and actions. Inconsistency leads to *cognitive dissonance* with the subsequent need to remove it. An example of their studies involved US and Spanish non-Muslim students. They were given accounts of mass violence committed by white Europeans and asked, 'How responsible do you think [white Europeans/you] are for the actions of [Darren Osborne/Anders Breivik/Istvan Csontos]'. After that they were asked: 'In general, how responsible do you think white Europeans are for the attacks of white supremacists?' In the second part, the procedure was repeated but with accounts of Muslim extremists and asked: 'Fatima Wahid is a Muslim woman who owns a bakery in southern France. How responsible do you think Fatima Wahid is for the Paris attacks of 2015?' Their results showed a reduction in collective blame of Muslims both one month and one year later, particularly among those with a strong preference for consistency in their thinking.

DO MACHINES THINK DIFFERENTLY?

Instead of attempting to address these prejudicial stereotypes through some form of intervention in the hope of reducing bias against specific human groups, we could instead bypass human judgement altogether and rely instead on technology to make our assessments for us. One could, in fact, look upon this as a kind of experiment testing the hypothesis that human beings' thinking is 'faulty' in some way – we have a defective way of making assumptions about people. One possible explanation

for this, as pointed out above, is that people base their judgements of other groups on what they see as specific attributes of the 'out-group' – 'they are not like us'. Whereas we would explain our own behaviour in terms of the context in which the behaviour took place ('I dropped the dish because I was in too much of a hurry'), we tend to often attribute the behaviour of others to some internal trait ('he dropped the dish because he is clumsy'). This is known as the *fundamental attribution error*. An extension of this leads to people assuming that all members of a particular group share the same attributes: 'women have no sense of direction', 'Americans are obsessed with guns', and 'people who support the Palestinians are anti-Semitic'. While this kind of error might lead to how the behaviour of particular ethnic, racial, religious, etc., groups are interpreted, there still needs to be some original source for the biased thinking in the first place.

An artificial intelligence (AI) system does not have the same cognitive system as human beings and should, therefore, be devoid of irrational prejudices and attributional biases and should rely instead on the raw data presented to it. If it is the case that humans cannot avoid biases, we should expect an AI system to make judgements free of such prejudice or bias. Typically, the form that many AI systems have used is based on a neural net where the system is trained on examples with feedback from any errors it makes. This kind of system has been used over many years in medicine. To pick but one example, doctors at Moorfields Eye Hospital in London trained an AI system to recommend treatments for 50 eye diseases. Whereas eye doctors were in agreement over 64% of the cases they viewed, the AI had 94% accuracy. If a computer system is capable of this kind of accuracy after training, the view was that an AI used, say, as a recruitment tool, would make better assessments of interview candidates than human assessors with their unconscious biases. We, therefore, have the basis of a potential experiment where the results of human assessors can be compared to the results of a computer-based system.

When Amazon piloted such a technological system as part of the recruitment process for new staff to avoid potential

gender and racial stereotyping, they found that the system was dismissing female candidates for no immediately obvious reason. On examination, what had happened was that the AI was trained on past recruitment decisions that were biased in favour of men. Hence, it continued to do the same thing. Technicians then adjusted the system so that it ignored words that specified a particular gender in CVs and resumés, such as 'woman' or 'male'. However, it continued to reject female applicants based on 'implicitly gendered words' such as 'captured' or 'executed' – words that male applicants apparently used more often than females. The AI turned out to be as biased as the humans.

There are, indeed, many examples of AI systems showing systematic biases in judgement and decision-making. In the US, parts of the justice system use AI algorithms to assess the likelihood of criminals to reoffend. One system, called the Correctional Offender Management Profiling for Alternative Sanctions was analyzed by Larson, Mattu, Kirchner, & Angwin (2016). They found that it incorrectly identified black defendants as being at higher risk of reoffending than white defendants, and white defendants as being of low risk.

Bolukbasi, Chang, Zou, Saligrama, & Kalai (2016) found that Google News articles contained sexist 'word embeddings' where words like 'receptionist' were linked to 'female'. They provide a list of 'extreme she occupations' and 'extreme he occupations' that allowed the generation of analogies called proportional analogies such as she:he:: queen:king (she is to he as queen is to king). This represents an appropriate analogy. They used a machine learning algorithm to generate the endings of proportional analogies relating to jobs – she:he::?:?. When the system ran, it generated not only appropriate analogies such as the one above but inappropriate ones he:-she::computer programmer:homemaker.

'The analogies generated from these embeddings spell out the bias implicit in the data on which they were trained. Hence, word embeddings may serve as a means to extract implicit gender associations from a large text corpus similar to how Implicit Association Tests […] detect automatic gender associations

possessed by people, which often do not align with self-reports' (Bolukbasi et al., 2016, p. 3).

Similarly, Kay, Matuszek, & Munson (2015) found gender biases in image searches for particular occupations. Also, in 2015, a search of images in Google for Chief Executive Officer (CEO) generated photographs of more male CEOs than female CEOs greater than the proportions in the general population, 11% vs 27% (Cossins, 2018). A cursory search by me for images of CEOs in mid 2019 showed a continuing similar disparity.

So, if humans have implicit biases, computer systems appear to have them too. The results of this virtual experiment show no differences between humans and other information processing systems (IPS). So, how do we now interpret the results given the original hypothesis that humans are somehow 'faulty' in their thinking? One view is that the biases in AI algorithms reflects human biases since the data they are trained on is biased. Hence, Richard Soccer on 17 January 2019 at the World Economic forum argued that 'Unmanaged AI is a mirror for human bias' (https://www.weforum.org/agenda/2019/01/ai-isn-t-dangerous-but-human-bias-is/). However, Karen Hao in MIT Technology Review (Hao, 2019b) points out that 'bias can creep in long before the data is collected' and at any stage: the coding stage, the collection stage, or the selection stage. In a later article she adds: 'Tech companies are built—and tech products are designed—with a "fantasy belief" that they exist independently of the sexism, racism, and societal context around them' (Hao, 2019a).

Thus, we have a chicken and egg situation. The biases come from somewhere but both AI systems and human beings are influenced by them. IPSs other than human beings are generating bias from the data they are trained on. Humans are also IPSs, so presumably we are doing the same thing. That is, our biases are due to being trained on the data through the environment and sociocultural influences we are exposed to. The environment, including the social environment, influences human activity and ways of thinking. At the same time, human activity and thought have an effect on the environment. We can therefore end up with a positive feedback loop whereby a particular opinion (or 'meme' in the original Dawkins (1976) sense), for example, is

noticed and shared which causes it to be noticed by more people and shared, and so on. This way a particular political or social view can gain currency, and we can track the rise of social movements as they increase in influence. These can be either positive or negative. Historically, there has been the rise of Nazism in Germany or Mao Zedong in China, the abolitionists in the UK and the US, the suffragette movement, white supremacist terrorism, the MeToo movement, and so on. These days such feedback loops are amplified by the 'echo chamber' provided by social media, but more importantly they show that aspects of our culture can be changed. Our thinking is the product of the culture we are immersed in. We learn about ethics, morality, technology, stereotypes, and all the rest whether we want to or not. According to Hinton (2017a, 2017b), if it is our determination to counter implicit sexism, racism, elitism, etc. and, given it is difficult to change our minds, we should concentrate on changing our culture since the bias is in the data.

SUMMARY

We often rely on stereotypes to make judgements about people from a particular social group since they are deemed to inherit the attributes of that group. Hasty induction can mean that the attributes of an individual from an unfamiliar group is assumed to be representative of the group. Such stereotypes can lead to potential biases against people's ethnicity, gender, accent, class, religion, university education, and so on.

The IAT has been used to elicit unconscious bias and diversity training has been established by firms and institutions to attempt to counter or make people aware of such implicit bias.

Stereotypes and associated biases are deemed by many to arise from normal categorization processes. According to some researchers, this can cause diversity training to be ineffective or short lived. A variety of bias reduction strategies may work by catering for individual differences among people. Collective blame is often ascribed to an out-group due to the actions of a small minority and can be reduced by pointing out the hypocrisy

and inconsistency involved, as collective blame is not directed at the in-group.

AI systems have been found to exhibit bias as they have been trained on data produced by people. It can be argued that people also are 'trained' on the same data. Rather than fixing the cognitive system, it would be more effective to change the culture that gives rise to the bias in the first place.

SUGGESTED FURTHER READING

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012, Nov). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. Retrieved from https://doi.org/10.1016/j.jesp.2012.06.003

Hinton, P. (2017). *Stereotypes and the construction of the social world*. London, UK: Routledge.

Payne, K., Niemi, L., & Doris, J. M. (2018). *How to think about "implicit bias" Scientific American*.