Finding unfairly biased assessments of people

Objectives and readers

This article is about the sensitive and difficult task of determining the extent of unfair bias in assessments of people, if any, in some commonly debated situations. Overall, the statistical issues are mostly simple and familiar but the conceptual issues need more attention and the research challenges are considerable.

The suggestions are applicable to all characteristics of people and all groups defined by their characteristics. There are no favourites.

Estimating the extent of unfair bias in assessments of people is not the same as trying to assess people fairly. (However, attempting a fair assessment and making comparisons is one approach that might be taken.)

The article was written for analysts such as statisticians and researchers but perhaps would be useful reading for politicians and voters too.

These are issues that generate a lot of heat in the news media, on social media, and even face to face among social groups and families. Emotive stories in the media about unfairly biased assessments of people are rarely supported by rigorous, reliable research properly analysed. Maybe there was unfair bias in an assessment, perhaps very serious, but often we cannot be sure. Academic work over decades has also been undermined by frequently made errors, usually attributing effects to unfairly biased assessment without adequately considering alternative explanations.

The analysis in this article is based on an understanding of fairness that captures typical views across the UK. This understanding is broadly consistent with truth seeking and enlightened self-interest, where good courses of action are effective for the whole society and for individuals (making a special effort to help stragglers).

It neither supports left wing nor right wing politics, but superficially may seem to be attacking the left. This is because complaining about unfair bias is now a characteristic tactic of the left and complaining about such complaints is characteristic of the right. Since this article is explaining how to find unfair bias reliably it also points out some common mistakes and they are made more often by the left than by the right. This is unavoidable. The only objective of this article is to promote more reliable identification of unfair bias.

If the political left were to apply the suggestions in this article their claims of unfair bias would be more accurate and far harder to debunk. They would be arguing on the basis of considerations that most voters can rationally agree with. They would avoid the collateral damage to those they seek to help that is caused by exaggerated or unfounded claims of unfair bias.

Similarly, if the political right were to apply the advice in this article they would debunk false claims of unfair bias more easily but also be forced to consider more deeply the practical consequences of unfair bias where it really exists.

The article begins by clarifying key terms, discussing why accurately finding unfair bias in assessments of people is important, and exploring some central issues. It then looks at methods for establishing bias and unfair bias in three common situations.

Key terms

Assessment of people

Typical reasons for **assessing people** are to decide whether to work with them, to lend them money, to decide how much to charge for insurance, or to decide how to treat them in prison. These assessments are not the same as the **decisions** they support.

A full assessment of a person might include an **overall assessment** and some **subsidiary assessments**. There should also be **evidence** that supports those assessments.

The overall assessment might just be a collection of subsidiary assessments with no explicit summary number, grade, or category. Alternatively, the overall assessment might be a number, grade, or category.

Numerical assessments may be expressed by **rating an attribute** of the person or as a **prediction** (perhaps a **probabilistic prediction**) of their future behaviour.

As an illustration of these ideas, consider assessing a person for a loan. Based on evidence such as their age, employment status, home ownership, credit rating from a rating agency, and reason for the loan it is common to calculate a probability of default (i.e. not paying back the loan and interest) and a loss given default. However, these two subsidiary assessments, focused on credit risk, are not the full picture. A person might be easy to serve because they are intelligent, educated, and able to follow rules and use a website without getting hopelessly lost. They may be commercially more interesting because they have the potential to become repeat customers for loans or other services. Consequently, the overall assessment of them as loan customers might include much more than their credit risk.

Assessment versus treatment

Unfair bias in **assessments** of people is different to unfair bias in **treatment** of people, though one usually gives rise to the other. Conversely, although treatment of people based on fair and accurate assessments of them will usually be fair treatment, in principle these are different and there may be exceptions.

You might accurately assess some people for a job but then give the job to someone who is not the best candidate because they paid you a bribe, have compromising photos of you, or support the same football club as you. That's an extra element of bias beyond the assessment.

Or, following the accurate assessments, you might give the job to someone who is not the best candidate as a kindness to that candidate, for some reason. Your accurate assessment allows you to know how much of a kindness you are doing.

Alternatively, society may wish to extend a kindness to some people, perhaps as compensation for a hard life, or because of a special contribution they have made, or for the economic reasons discussed later in this article. Rather than expect particular employers to bear the cost of that kindness the government can provide compensation to employers that they can then include in their assessments of people. The employer then simply makes the fairest, most accurate assessment of job candidates that they can, including the effects of government compensation (e.g. for pregnancy or disability), and selects the best candidate as usual.

This way the government knows the cost of its kindnesses.

In general it is better to make assessments that are accurate and unbiased even if the action that is expected to follow is not directly based on those assessments. This is for three reasons:

- Accepting inaccurate, systematically biased assessments as true and accurate could lead to poor decisions on other matters. (This includes the people advantaged by biased assessments making less effort to improve themselves than they should.)
- It is easier for people disadvantaged by society's kindness to others to accept that a special kindness is being given than to accept as accurate an assessment that is clearly incorrect.
- It is important to know how much special kindness is being given to avoid giving more than is fair to others.

If you want to help someone because you think they suffer some kind of disadvantage then you need to understand exactly how great it is. If you do not know then how can you decide what level of help is fair and how can you justify providing special kindnesses?

E.g. Suppose you think that mathematicians get a raw deal in employment interviews because people expect them to have poor social skills. To tackle this you might suggest requiring employers to ensure that their ratings of social skills for candidates are, on average, equal for mathematicians and non-mathematicians. But what if, on average, mathematicians do have different social skills to non-mathematicians? Hypothetically, what if poor social skills really are more common among mathematicians?

If your idea leads to people rating mathematicians the same when they are not you will struggle to justify effort to identify young mathematicians with poor social skills and offer them help. You cannot argue that they are the same as everyone else but still need special help.

Several methods developed to reduce unfair bias in machine learning involve pushing the analysis away from objective reality in order to produce an assessment that, if acted on directly, would be considered fair. Typically, if there are two groups defined by a protected characteristic and one, in reality, on average is less capable than the other in some task then the algorithm tries to eliminate evidence of that real difference. This is not a good approach.

Having said that actions might not be consistent with assessments of people for some good reasons, two final points should be should be made.

First, the reason for acting inconsistently with an objective assessment should be clear and sound. It may be because of a known limitation in the achievable assessments, for example. Such interventions should be stopped if experience shows they do not work.

Second, there is a better approach to the common challenge of trying to improve the productivity of people who have had a raw deal in the past but have the potential to thrive in improved circumstances. That better approach is to assess a person's future potential given the support that could be made available. The assessment should be based on all available relevant information, not just one or two special characteristics.

This means the assessments are improved and actions once again become consistent with assessments.

Bias

Bias is measurement error that is not random, though it need not be deliberate.

The bias might be identifiable statistically because the average of a sequence of measurements of the same thing is too high or too low. Alternatively, it might be that the exact mechanism driving the bias can be identified, such as the effect of temperature on a metal measuring rod.

Fair and unfair bias

This paper will distinguish between **fair** and **unfair bias**. Not everyone makes this distinction and that can lead to unkind treatment of people who do not deserve it.

Fair bias is bias that occurs despite honest, diligent efforts to measure accurately and without bias. It is still bias but it is not the sort of bias for which people should be scolded or punished. Being fairly biased does not make you a bad person.

Someone whose measurements are fairly biased will usually be quite willing to make corrections if the bias is identified and there is a practical way to do better. They just need improved ideas and information.

In contrast, unfair bias is bias caused by vested interests, laziness, or irrationality driven by emotional issues. Scolding or punishment may be appropriate. Being unfairly biased does make you a bad person (at least on that point and at that time).

Unfair bias may also respond to improved ideas and information (partly because it is harder to get away with unfair bias when information improves) but it may be appropriate to scold or punish those who have been unfairly biased. They may need more motivation to behave better. This may be to get them to make more effort, to overcome vested interests, or to push through irrational, emotional issues.

To illustrate these ideas, imagine that some physicists need to make an estimate of a physical constant that is hard to measure.

Accurate: After years of work they devise a gadget that can make the estimate very accurately and consistently. Their estimate is now so accurate it is misleading to call it an estimate. It's a measurement.

Biased: Before that breakthrough their estimates varied quite a lot, partly because different people used different methods at different times. Due to different biases some methods tended to give under-estimates while others tended to give over-estimates.

Fair bias: However, the bias might not have been unfair. Indeed, with physicists estimating a physical constant it is unlikely that they would be subject to unfair bias. They will usually seek accuracy and their errors will be the result of honest errors made as they try hard to do something difficult.

Such bias is fair bias.

Unfair bias: But suppose instead that the hypothetical physicists belonged to two competing research groups. For years one group had been developing and promoting a theory that predicts a particular value for the physical constant. The rival group had a different theory that predicts a slightly different value for the physical constant.

Now there is a risk of unfair bias. Perhaps the groups would have preferred methods and corrections that nudge the estimates towards the values they predicted from their theory.

Related terms

Several specific terms are related to unfairly biased assessments of people, including: racism, sexism, homophobia, Islamophobia, xenophobia, transphobia, bigotry, prejudice, stigma, and hate.

Although most people would probably say that these refer to unfair bias, and this is the way that UK law usually understands them, some have redefined these terms in more expansive ways. For example, some have argued that all men are sexist against women, even if they deny it, because they are born into a patriarchy that oppresses women. On this basis it is sometimes argued that women simply cannot be sexist against men, no matter how unfairly they talk about or treat men.

This article will not discuss these arguments over definitions. Instead, it uses the term 'unfairly biased assessment', which is more precise, more self-explanatory, and relatively free of emotional associations and politics. However, occasionally examples are used where other terms were used by the researchers involved.

In writing about these issues it is common for authors to make no distinction between fair bias and unfair bias. They also often make no distinction between fair and unfair discrimination (i.e. treatment). They just write about 'discrimination' as if it is always a bad

thing. In fact we discriminate frequently and fairly. Life would be much harder if we did not. For example, we prefer to undergo surgery with a surgeon who has skill, knowledge, and experience. Competence and honesty are some common reasons for discriminating fairly.

At stake

Claims about unfairly biased assessments of people should be accurate because there are significant negative consequences from both under- and over-estimating the extent of unfair bias. It is not appropriate to err in one direction in the belief that the consequences of error in that direction are less than the other.

We want to prove unfair bias convincingly when it exists, but avoid making unwarranted allegations when it does not.

Today in the UK, failures to do both of these are common. Countless incidents of unfairly biased assessments go unchallenged while the news media frequently feature stories that exaggerate the extent of unfair bias. Both mistakes are harmful.

Making objectivity harder, there are people who would prefer the extent of unfair bias in some assessments to be under-stated, and people who would prefer it to be over-stated.

When people, such as politicians and 'activists', are trying to promote the interests of their favoured groups, some will exaggerate the extent of unfairly biased assessments of those groups in the hope of getting more concessions to help them. Others will try to down-play the extent of unfair bias in the hope of reducing the concessions that have to be made.

Some people will do this as far as they can get away with it and may feel that they are morally justified in doing so. They want good things for a group they think has had a raw deal, or would get a raw deal, and if a bit of exaggeration or down-playing might help then they see that as acceptable, even admirable.

The following tactics may be used in this battle of rhetoric:

- Distorting the facts, or just making some up.
- Redefining the words used (e.g. 'racism', 'sexism') so that innocent behaviour can be labelled as if it is evil, or vice versa.
- Generalising the accusation or denial to include everyone in a large group.

These tactics backfire when decision-makers become suspicious, potentially leading to delays, the end of discussions, and actions that are prompted by a desire to show they will not be tricked or bullied.

In the following explanations the main actors are the alleged victims of bias, their alleged oppressors, and their alleged defenders (defending the victims from oppression without themselves being victims). The word 'alleged' is used because this analysis is only concerned with situations where perceived unfairness is very different from actual unfairness.

The harms of over- and under-estimates of unfairness affect alleged victims, alleged oppressors, and our wider society. One reason for this is that our societies need all the productive people they can get. There is no shortage of useful work to be done thanks to an aging population and the urgent need to make our way of life sustainable. Although there are people who want paid jobs but cannot get them, this is because there is inefficient matching and some people are very hard to make productive. The fact that important work goes undone year after

year is evidence that shortage of ability is a large problem.

The detail provided below on under-estimation is less than for over-estimation because the harms from under-estimation are probably better understood by most people already. This is not an argument that the consequences of over-estimation are greater than of under-estimation.

Harms from under-estimating unfair bias

If unfairly biased assessment continues undetected and uncorrected then harm is done through a number of mechanisms. The alleged victims are not the only ones to be harmed.

Unfairly biased assessments can lead to pointless, stressful, harmful conflict, up to and including beatings and death.

> E.g. Violence might arise from overestimating the risk of violence by a person and defending against it unnecessarily.

Opportunities to develop people may be missed. This harms those people overlooked but also the whole of society because their abilities are not developed.

> E.g. Children from families where parents do not support academic achievement may struggle when young and never recover. And yet, with the right encouragement, they would have blossomed later.

People may be matched inefficiently to work roles. This harms the whole society, but especially those who miss out on suitable roles and those who rely on the roles poorly filled.

> E.g. Some highly productive workers might be barred from roles for reasons that have nothing to do with their suitability for the roles.

> E.g. Workers with poor experience but good potential might be overlooked by an employer who fails to assess their ability to improve their skills and productivity.

People may wrongly attribute their own lack of progress to lack of ability and effort when in reality they are being discriminated against unfairly.

> E.g. A researcher with a good but unconventional idea for a medical treatment might conclude there must be something wrong with the idea when in reality the problem is narrow-minded colleagues and committees.

> E.g. A child may think that they are bad at mathematics because their mathematics teacher is always angry at them. The reality might be that the teacher hates their whole family for religious reasons.

Opportunities to cooperate gainfully may be missed. Friendships and romances that would have worked might not get started.

Low expectations might lead to lower performance.

> E.g. Assessing a child as inherently weak at mathematics might lead to them being taught too slowly, covering fewer topics, and not learning techniques that promote higher performance. The low assessment might be due to failing to take into consideration the effect of two earlier years of poor teaching. In the UK this sometimes happens in primary schools and in preparation for GCSE examinations.

The importance of low expectations is not clear. Although this mechanism has sometimes been seen as hugely influential, the scientific evidence for this is disputed and not convincing. If those low expectations lead to different educational decisions then, of course,

there is an effect. However, if it is just some slightly discouraging remarks or body language from time to time it is unclear what effect, if any, this has. For some students it may even spur an extra effort to prove people wrong.

Harms from over-estimating unfair bias

It is also the case that exaggerated perceptions of unfairly biased assessments are harmful. They can affect attitudes, lead to unjustified or exaggerated accusations of unfair bias, lead to poor decisions, and even to unfair 'reverse' discrimination. The harms affect alleged victims and oppressors, and the wider society.

As these harms are perhaps less well known the following paragraphs explain them in detail.

Alleged victims who over-estimate the bias they face may be mistakenly upset at the thought of (imaginary) future oppression by society or powerful people.

> E.g. People may react angrily online because of thinking that large numbers of people are unfairly biased.
>
> E.g. People reacted in different ways to the election of Donald Trump as President of the USA and we can be sure that at least some people over-estimated his unfair biases and felt threatened unnecessarily as a result.

They may go further and behave aggressively against people they imagine to be oppressors with evil intentions.

> E.g. If they think that law enforcement officers are evil oppressors then they may react resentfully or aggressively towards them and so get into legal trouble that otherwise would have been avoided.
>
> E.g. If they think that people in positions of authority (e.g. teachers, managers, law enforcement officers) are unfair oppressors then they may be less willing to comply with instructions or apologise for things they have done wrong, leading to problems that should have been avoided.
>
> E.g. If they think that most of society is unfairly oppressing them then they may think they are morally justified in breaking some laws. They may be more willing to break the law as a result, or may be more likely to think that a jury will find them innocent, even if they have broken the law.

Alleged victims may avoid choices that would have been good for them due to unnecessary fear of bad reactions or unfair discrimination and subsequent failure.

> E.g. People who think a particular type of job is 'dominated' by another type of person might avoid that job and so miss out on a good career, perpetuating inefficient allocation of labour.
>
> E.g. Imagining an employer to be an evil oppressor might lead to rejecting a good job offer.
>
> E.g. A person might stay in a racial/ethnic ghetto or a gang due to imagining that racism would make success outside impossible.
>
> E.g. A person might worry about 'coming out' because of mistakenly expecting a big negative reaction.
>
> E.g. People might avoid pleasant holiday locations because they wrongly think they will be 'hated' there.
>
> E.g. A woman might fail to report a genuine rape due to thinking the police are mostly sexist and will not act.
>
> E.g. A community might fail to cooperate with the police leading to

failure to control crime affecting that community.

Alleged victims may put less effort into activities than they should, expecting their efforts to be wasted due to unfair discrimination and subsequent failure or rejection.

E.g. A person might give up too soon when school work gets difficult, thinking the extra effort will be wasted. They may believe that, even if they strive and get slightly better qualifications, they will be discriminated against and rejected from universities and jobs in future anyway.

E.g. Later in life a person may put a low level of effort into a job, thinking that doing more will not lead to higher pay or promotion because of unfair discrimination.

Alleged victims may misinterpret innocent actions by alleged oppressors as being driven by unfair bias, and so over-react or assume problems are caused by insoluble bias when they are not.

E.g. A person might react angrily to their spouse because they thought an action was based on sexism when it was not.

E.g. As jury members, people may be too willing to believe that a crime was motivated or aggravated by 'hate'.

E.g. As journalists they may be too quick to assume that an incident was motivated by unfair bias, creating stories that perpetuate exaggerated perceptions of bias among some people.

They may react to people within an alleged oppressor group in an inappropriate way that prevents warm relationships forming and prompts reactions that seem to show the expected unfair bias.

E.g. Anticipating rejection, people might not try to make new friends in another demographic group.

E.g. They might not start a romance because of assumed discrimination by others.

E.g. They might not try to join a club because of imagined prejudice[1].

E.g. They may be angry and resentful towards innocent individuals due to a negative attitude towards a whole group that they believe to be unfairly biased oppressors.

E.g. According to Bergsieker et al (2010), when a white person expects a black person they meet to be unfriendly and the black person expects the white person to think them stupid, the white person tries to be friendly and informal while the black person tries to be more formal and reserved. This leads to an uncomfortable encounter.

Alleged victims, faced with evidence of problems that they have a role in solving, may fail to act, thinking that the problem does not really exist and has just been made up by alleged oppressors, or thinking that the problem is solely caused by the alleged oppressors.

E.g. People might not bother to speak clearly because they think their strong accent is a part of their culture that should be respected rather than a weak skill that should be corrected. Objectively, speaking more clearly is a good thing.

[1] Some evidence from psychological research (Maltese et al, 2016) suggests that people who are very sensitive to unfairness against themselves, and quick to perceive it, tend to be anti-social and uncooperative to avoid becoming victims.

E.g. A person who thinks their facial tattoos and piercings should not hurt their job prospects because that would be unfair discrimination needs to understand that their choices signal a variety of judgement and character flaws, even if we exclude consideration of how people react to a person who looks like they do. The treatments are expensive, painful, risky, and signal vanity, or insecurity, or lack of impulse control, or all of these.

E.g. A person who thinks that employers should ignore their morbid obesity because considering it is unfair discrimination and wrong. Being less effective at work is just another of the negative consequences of being greatly overweight.

E.g. Some may fail to address a solvable problem with their spouse because they wrongly thought the spouse's behaviour was the result of their sex (and therefore an inherent characteristic).

E.g. A person might fail to address a problem in their approach to the opposite sex, having generalised from some bad experiences. ('Radical feminists' and 'MGTOW guys' often seem to have a bitterness that I suspect has come from a bad relationship, or a string of them perhaps.)

If alleged victims combine failure to solve their own problems with frequent complaints that their problems are caused by (innocent) alleged oppressors then this may lead others, not just alleged oppressors, to form a negative view of the alleged victims. It is worse if the alleged victims claim that their alleged oppressors are bad people and if these claims are repeated and widespread.

Innocent alleged oppressors who have seen others attacked with false or greatly exaggerated accusations of unfair bias may avoid discussing and solving problems that affect many, including the alleged victims.

E.g. They might not discuss and work out solutions to problems because the whole subject has been made sensitive by frequent verbal attacks alleging unfair bias (e.g. problems arising from large scale immigration).

E.g. They might allow bad ideas and actions to pass without criticism or control through fear of being called 'racist', 'sexist', or something similar.

E.g. Police might not enforce the law in some situations due to fears of being seen as racist, sexist, political, or otherwise unfairly biased.

E.g. Social workers might not investigate odd behaviour in a family because they think it might just be 'cultural'.

E.g. People may avoid addressing bad behaviour of a group through persuasion, education, or incentives because this could be seen as unfair discrimination.

E.g. Researchers might avoid performing scientific research on some topics because they are too sensitive (i.e. too often lead to verbal attacks alleging unfairly biased assessments).

E.g. People might not speak out to make reasonable contributions in a group discussion for fear of being seen as unfairly biased, or told their contribution is worthless because of their demographic group membership(s).

When they do act it may be inefficiently due to fear that efficient solutions will be attacked as unfair.

E.g. Law enforcers might give crime by black people against other black

people less focus than police crime against black people, even if the former is much, much more common.

E.g. Anti-terrorism forces might not manage risk efficiently for fear of being seen as unfairly biased against people who statistically do pose a greater risk.

The same fear may discourage them from innocent participation in society.

E.g. A person might avoid being 'patriotic' (i.e. supporting their country) for fear of being seen as racist or xenophobic.

E.g. A person might not join a club for fear of being seen as unfairly biased.

Another effect of frequent, overblown accusations against members of a group may be resentment by many innocent members of that group. This may create negative feelings that seem to be evidence of the negative bias alleged.

E.g. Persistently insinuating or alleging much higher levels of unfair bias within a large demographic group than is really present is annoying to members of that group who are not unfairly biased.

E.g. Arguing that all people in a large demographic group are unfairly biased whether they know it or not is particularly annoying to people in that group who are not unfairly biased by any widely acceptable definition.

False or exaggerated accusations of bias may be so common that genuine accusations are not recognized or acted on, or the reaction may be delayed or limited. This is the Cry Wolf effect.

Further harms arise when false accusations lead to action to address the non-existent or greatly exaggerated problem. Resources may be wasted trying to solve a non-existent problem that should have been used to solve real problems.

E.g. A comfortable living may be given to people (usually academics and writers) to raise issues exaggerating unfair bias – work which is worse than useless.

E.g. Charities may expend effort and funds on tackling non-existent or negligible unfair bias.

E.g. News reports may spend time on non-existent unfair bias.

Interventions may be implemented that are not needed or are excessive and create unfair bias in favour of the alleged victims and against the alleged oppressors.

E.g. A baseless intervention might set easier course entrance requirements for a particular type of student, even though they then too often struggle to keep up.

E.g. Recruitment quotas might lead to not hiring the best person for the job, at the expense of all other stakeholders, not just the better candidates who otherwise would have got the jobs.

These can lead those privileged and others to wonder if their success is truly deserved.

Good people may be removed from roles because of false accusations of bias.

Defenders of alleged victims may take advantage of feelings of victimhood for their own ends, even if they genuinely believe the oppression is real. They may exaggerate it and repeat accusations as often and as loudly as they can.

E.g. People may be recruited into criminal gangs, religions, and terror groups in part through developing and exploiting an exaggerated sense of victimhood.

E.g. People may vote for the wrong political party or give support to the wrong political groups because of exaggerated perceptions of unfair discrimination, perhaps collusive, and the belief that the party/group will fight it for them.

Overall, a society may become less harmonious and cooperative due to increased perceptions of victimhood, generalised from resentment of individuals to resentment of whole demographic groups, due to this being encouraged by alleged defenders (including political groups and journalists).

Some central issues

Sensitive characteristics

In some countries, including the UK, the law has a list of 'protected characteristics' that are subject to rules designed to reduce 'discrimination'. The definitions are provided in the law. The protected characteristics in the UK's Equality Act 2010 are age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex, and sexual orientation. Treating people differently because of one or more of their protected characteristics is limited by the law.

If a characteristic is protected then that does not mean that it has no practical significance. For example, people with some disabilities are, on average, less capable of doing some tasks than people without those disabilities. If ability was not affected I'm not sure you could claim to have a disability. Similarly, women in the UK are, on average, longer living than men, shorter in stature, and have different medical risks.

Also, women in the UK today are more often interested in occupations that involve working with people while men are more likely to prefer working with machines. Whether this is in any way driven by biology or purely the result of society's teaching is controversial.

Skin colour per se is often regarded as something that should not need consideration because it makes no difference. And yet a darker skin does confer the advantage of tolerating stronger sun and creates less vitamin D than pale skin, a disadvantage in colder, darker climates.

The practical significance of characteristics depends on context. Something that is relevant to performance in one type of job may be irrelevant to another.

What the protected characteristics have in common is that there are people who don't want to be treated badly because of the characteristic and who think they have been or would be.

The reasons for this differ. For example, with race the most common reason for suggesting protection is that, for historical reasons, some races in some countries are relatively behind in their education and career progress. The idea is that, if they are given a helping hand, they can gain more opportunities and gradually close the gaps. In future they will not need that extra help.

In contrast, for disability the idea is to give disabled people continuing help so that they can thrive in education and careers despite disabilities that in many cases put them at a genuine disadvantage that will not completely go away over time.

Then, with sexual orientation, the focus is more often on reducing mean treatment of people who are not heterosexual.

There are other characteristics that could be the basis of unfair assessments and discrimination but currently are not 'protected'. These include birthday

(relative to the school cohort cut-off) height, handedness, hair colour, physical attractiveness, voice pitch and quality, extroversion/introversion, home town, educational establishments attended, former employers, sports teams supported, club membership, parental wealth, parental education, family connections, inherited wealth, occupation, fashion sense, tattoos and body piercings, hyperhidrosis (a condition that leads to permanently sweaty hands), and the way you pronounce 'scone'.

As with protected characteristics their relevance depends on context.

If we want to avoid unfair bias in assessments of people then it is important to consider all potential causes of unfair bias, not just legally protected characteristics. The number of possible causes is too numerous to list. This leads to the idea of assessing each person as an individual rather than as a member of one group.

Characteristics of people that might be the basis of unfair assessments or treatment will be called 'sensitive characteristics' in this article.

The conclusions reached in this article might not be consistent with the law in all countries. Some laws are wrong. Focusing on just some characteristics that may be the basis for unfair bias is just one example.

Motivation

The purpose of identifying unfair assessments of people is to promote fair assessments and, ultimately, fair treatment of people.

Various reasons for fair treatment have been put forward but two reasons that people from all perspectives should be able to agree on and support are these:

- Fair treatment of people in education and work promotes a society that is, overall, more productive and efficient. Waste of human resources is reduced.
- Some work groups are more effective if there is the right amount of the right type of diversity among their members.

Realising labour's potential

Overall, it is good for society if as many people as possible fulfil their potential as positive contributors. This might be through paid employment, unpaid services, or simply being easy to look after (as with a well-behaved child) and not disruptive. In contrast, if many people idle at home, going out only to make trouble and damage property, and living on government handouts, society will struggle. We all should try to do our bit.

If there are groups who have a historical disadvantage that holds them back in education and work then interventions of some kind may help them to catch up and improve their contribution to society, learning more and doing more that is useful. This may be true even if their present disadvantage manifests itself in what appears to be bad behaviour and poor choices. The interventions may take more than one generation to be fully effective but in the end it should be worth it for everyone.

If there are people with personal disadvantages or special characteristics that are not historical then interventions of some kind may still help them to increase their contribution within their lifetimes.

If there are people who have been denied opportunities simply because they are unusual then interventions may help them too.

Because there are many factors that lead to some people getting better

opportunities than others it is important to consider all relevant factors for each individual.

> E.g. Imagine that two people are being assessed for their potential as athletes. One person grew up in a poor family in a rough neighbourhood but took to going to a free local gym run by a charity and made friends there. The other person grew up in a rich family in a wealthy neighbourhood but was too busy with school work to do more than some jogging. Now imagine that both currently have equal physical abilities and equal motivation. Who has greater potential as an athlete?

There are limits to how effective interventions can be. We are each born with potential encoded in our DNA, and that provides advantages and disadvantages. Disability is not disability unless it limits our ability in some way. People without a womb cannot give birth to children.

Disadvantages that are not biological may still be so profound that they cannot be entirely eliminated within one lifetime.

> E.g. A person whose early education in mathematics was poor may never overcome it. They may remain fearful and incompetent even though, with a great effort on their part and just the right help, they might have become competent mathematicians.
>
> E.g. An immigrant to a country who is not a native speaker of that country's language will have a disadvantage in most activities as a result. That disadvantage may last for the rest of their lives and even affect their children, who grow up in a household with imperfect language skills.

Some disadvantages are the result of decisions the person took, freely, that simply turned out to be unlucky choices.

For all these reasons, some people are very, very hard to help.

Relevant diversity

Teams can sometimes be more effective if their members have varied strengths and weaknesses, varied knowledge and experiences, and varied contacts. This is not always the case. Sometimes differences can lead to conflict. Sometimes every team member has the same task and it is best if everyone is expert at that task.

Where diversity can help at all it must be the right type of diversity. For example, if a team is searching for a treatment for a virus then it is not helpful to gather members from a wide range of religions. In contrast, if the team's job is to promote some kind of cooperation between different religions then a diversity of beliefs could be useful. (But, even here, it might be better to rely on someone who is not a follower of a religion but has studied religions in detail. This might provide information and insight with less risk of conflict.)

Very often diversity on protected characteristics is not relevant to the purposes of a group, but if the net is cast wider to consider all sensitive variables there might be something that is more relevant.

Where diversity might help it does not necessarily have to be achieved by matching the composition of the work team to the composition of the general population. There is usually a tension between representation and ability. Usually we would like decisions to be taken by people who are unusually intelligent, wise, knowledgeable, open-minded, and able to weigh fairly the legitimate interests of all stakeholders. People with those qualities rarely have deep personal experience of the life problems that most people have because

their smart choices have allowed them to avoid or solve most of those problems.

Incentives

Whatever is done to promote fair assessments and treatment, societies still need incentives. Most people need to know that they will be materially better off if they make a useful contribution instead of just relaxing on the sofa all day.

Most people do jobs that they would not do without pay because the jobs are tiring, boring, uncomfortable, stressful, involve very long hours, a long commute, or working with people who are sometimes mean to them. We pay to meet people at a club and have a good time. We usually have to be paid to meet people at work and do something useful for others.

Parents who look after their own children, people who take care of their own homes, people who tend their own gardens, often do it partly because they like it. Still, even here there are times when it would be nice to rest and let someone else take over for a while. The cost of paying someone to do that for us helps keep us making an effort.

In countries with an economy that makes good use of fair markets, organizations looking for people to fill roles are able to compete with each other through pay and conditions to gain the services of productive employees. This tends to mean that the roles that have high economic impact on society are also highly paid, attracting the most effective people to the roles where they can do most good. Overall, this is good for everyone.

The end goal

We want people to find roles in which they are useful to others, with labour efficiently distributed for everyone's benefit and relevant diversity in place where it will help.

This means that people will be paid different amounts. Even in a perfectly fair society, calculate average pay for every possible grouping of people based on sensitive characteristics, every intersection, and every organizational unit and there will be differences on average. This will be partly by chance and partly because some groups really are less interested or productive than others, on average.

Similarly, calculate the distribution of roles by every sensitive characteristic, every intersection, and every organizational unit and there will be differences compared to the overall population. These too will be by chance and because some groups really are less suited or less interested than others, on average.

Eliminating differences in average pay between groups and making every group a representative sample of society are not goals that should be of over-riding importance. They may be indicators of some kind of positive progress but as primary goals they can be damaging.

For example, one way to eliminate differences in average pay between groups would be to pay everyone the same amount. However, this would eliminate incentives and produce an unproductive society.

Another approach that would achieve equal average pay and representative groups would be to consider only averages over very large organizations and allocate people to roles randomly. However, who does what is not just a decision for the employer or a government. Some people will not want to do the jobs randomly allocated to them. This scheme would produce an unproductive society with many deeply unhappy people.

Still another approach would be to use an algorithm that computes an overall score

reflecting how close a workforce gets to equal pay for groups and representation of the overall population, and then use that to guide recruiting decisions. If some roles proved hard to fill with just the right sort of person then those people would have to be pushed to take the jobs. This scheme too would produce an unproductive and unhappy society.

The end goal should be to get everyone to be **as productive as they can be** (taken across all their contributions, not just paid work), measured over more than one generation if necessary. This is not the same as getting everyone to be equally productive.

Genetic potential

One distinction that is often crucial in assessments is that between a person's genetic potential and their actual condition now. For example, imagine a person who has the genetic potential to be a leading intellectual but is poorly nourished as a child, given no education at all, and starts work from childhood gathering food in a war-torn jungle.

Even when we think that two groups of people should have, on average, the same genetic potential there will usually be differences in their actual abilities and achievements due to circumstances.

Some of those differences will never be eliminated during their lifetimes, even by the most supportive care. It is too late.

We also cannot assume that two groups of people will have the same genetic potential simply because they are the same species. Selection pressures may have operated within their lifetimes. For example, if migrating to a different country from a failing country involves complying with complex administrative procedures and paying a considerable amount of money then successful migrants may be more capable, on average, than those they leave behind[2]. That may be partly for genetic reasons.

Genetic potential also, typically, is partly inherited. This means we have to accept that, for example, Ethiopians, on average, are better suited to long distance running than Nigerians, because of their genes. The children of famous intellectuals will, on average, be smarter than those of less brainy parents, because of their genes. The children of good looking parents tend to be better looking than most people, because of their genes.

Great training and education can overcome genetic differences in individual cases, provided the training and education are not given equally to everyone. However, this does not eliminate the underlying genetic differences.

Onus of proof

Debates about assessments that are, potentially, unfairly biased sometimes revolve around who has to 'prove' that bias occurred. Sometimes one side will say that some evidence they have found looks at least fishy and then say that it is for the other side to prove that no unfair bias was at work. Sometimes one side will say that no bias has been proven beyond doubt and therefore there is no evidence of bias. Both positions are wrong.

What should happen is that the probability of each possible hypothesis (e.g. biased versus not biased) being true is adjusted as evidence is considered. At any time we will have a probability of truth attached to each possibility. Typically, that means we neither think there has been unfair bias nor that there has not, though one will often seem more likely than the other.

[2] This is just one theoretical possibility. For a particular example of migration, migrants might be less capable on average, or perhaps the same as those they leave behind. The point is that some selection pressure may be involved.

This kind of uncertainty is normal, natural, and appropriate. To conclude definitely too early would be a mistake.

When it comes to public accusations of unfair bias the usual approach is 'innocent until proven guilty'. Accusations of unfair bias can be extremely damaging, even when untrue, so great care should be taken to make them only when there is very strong evidence behind them.

> E.g. Does a high proportion of men, or a high proportion of women, in a particular occupation provide strong evidence that the less frequent sex is being unfairly assessed for jobs or suffering some other form of bias? According to Careersmart, in the UK in 2019 about 98% of nursery nurses and their assistants were female. Almost 99% of welders were male. From that almost nothing about unfair bias can be deduced. It certainly is not safe to conclude that these differences are entirely due to unfair discrimination by recruiters.

Prior probabilities

One possible reason for arguments over onus of proof lies in the logic of using evidence. The Bayesian approach gives us the insight that we start with prior beliefs then learn from evidence and revise those to form posterior beliefs. One consequence of the logic of belief revision is that if we believe something with complete certainty then no evidence will shift our view.

A person who is convinced that assessments are always unfair can be confronted with evidence that is *almost* entirely inconsistent with unfair bias and yet still continue to believe that unfair bias has happened again. The converse is true for people utterly convinced that unfair assessment does not happen.

In contrast a person with some doubt is dramatically more responsive to evidence.

Automated assessments

Today many assessments of people are carried out by computer. The main types are these:

- Performance statistics calculated very simply from records of work (e.g. time utilisation) or tests (e.g. a score on a multiple-choice exam marked by computer).
- Assessments computed by software that encodes human expertise as explicit rules, formulae, and parameter values put in place by a human author.
- Assessments computed using a statistical model of past assessments made (trained using data of past cases and the assessments they received).
- Assessments computed using a statistical model of past performance (trained using data of past cases and the performance those people achieved in reality).

Where a statistical model is used there are two main cases:

- The model contains weights or other parameters that give us a good idea of how the model is using evidence to reach its assessment (e.g. linear regression).
- The model is, effectively, a black box that uses the past experience but does not explain itself (e.g. a typical neural network).

Automation does not necessarily eliminate bias, or even unfair bias, but it does offer the potential for reducing bias of all kinds compared to assessments made by unaided humans.

The machine does not get tired or make mistakes. It is consistent, which is one reason why simply automating existing

assessments usually improves predictions. The algorithm does not respond to special pleading or bribes.

It can also be easier to examine automated assessments for evidence of bias. We can see the evidence used, the rules applied, and often there is a convenient electronic record of the inputs and results that can itself be studied using a computer.

Automation using a computer requires a more explicit, intellectual approach from people and that is also helpful. Recent work looking at possible bias in 'machine learning' from 'big data' has clarified issues that also apply to human assessments but have not previously been tackled as clearly.

When possible unfair bias is to be assessed there will often be more evidence to work with (see Table 1).

Things that might be available:
• Training data
• Model (structure, parameters)
• Model fitting process/rule
• Test data
• Assessments produced
• Actual performance of people
• Alternative training data
• Alternative test data

Table 1. Evidence that may be available when searching for bias in automated assessments.

Decisions to cooperate

Having clarified what unfair bias in assessments of people is and why it needs to be identified accurately it is time to look at some common situations where unfair bias in assessments is studied. The first is decisions to cooperate.

In many situations we have to decide who we will cooperate with. For example, who would we lend money or other assets to, or go into business with, or do a simple favour for? If we seem more willing to cooperate with some people than others, is that the result of unfairly biased assessments?

The mere fact of preferring to cooperate with some people and not with others is not in itself enough to demonstrate unfairly biased assessment. There are some legitimate reasons for preferences and honest errors that are not unfair.

The opportunity problem

A frequent concern is that people might over-react to something about a person that is seen as negative.

> E.g. Imagine a person who was convicted of theft while young and now seeks opportunities such as education, employment, and volunteering positions. If people with those opportunities to give over-react to the criminal record then the person may get no opportunities to progress and show that they have changed for the better. It is reasonable for people to be cautious and less willing to give opportunities because of the criminal record, but this can be taken too far.

The problem is not that the person is assessed lower than they would be without the negative characteristic. This is objective and reasonable.

The problem occurs when they are assessed in a lazy, over-simplified way that completely excludes the person on the basis of one relevant but inconclusive piece of evidence. This is an over-reaction.

Instead of a downward spiral created by lack of opportunities such a person needs a succession of controlled opportunities in which they can show they are trustworthy. This may mean that a person who otherwise would not have been trusted to do useful work becomes an economic contributor to society instead of just a drain[3].

They have a better chance of getting these opportunities if:

- there is a shortage of willing labour;
- they are willing to accept low pay in exchange for an opportunity;
- the employer is compensated by society; and
- the employer's assessment is based on a wider pool of relevant information and the negative feature does not lead to automatic rejection.

Unfairly biased assessment

More generally, an assessment is unfairly biased if it is objectively wrong due to laziness, incompetence, stubbornness, emotional irrationality, or selfishness. An assessment that is wrong despite an honest attempt to gain and use relevant information rationally is not *unfairly* biased. Signs of unfair bias in assessments include the following:

- Considering and responding to evidence that is plainly irrelevant to the assessment, both directly and statistically.
- Continuing to consider and respond to evidence that is clearly no longer relevant to the assessment because of new information.
- Using as evidence a data set that is obviously affected by a bias (e.g. a selection bias when sampling, a training set of past assessments that were themselves clearly biased) or continuing to rely on such data even when the bias is obvious.
- Reaching an assessment that is systematically biased with respect to the evidence in some way that should be obvious.
- Reaching an overall assessment that uses subsidiary assessments that are plainly irrelevant.
- Reaching an overall assessment that plainly fails to capture important characteristics of the person relevant to the purpose of the assessment.
- Failing to make a reasonable effort to gain further, relevant evidence.
- Using a characterisation of the members of a group that is itself unfairly biased, having the wrong average, the wrong spread, or both.
- Failing to adjust an assessment in the face of new, relevant evidence.

Several of these are explained further below.

Irrelevance: What is relevant may depend on what type of cooperation is considered.

> E.g. A person's race and sex in themselves may not be directly relevant to their trustworthiness for a loan but would be directly relevant if you were looking for someone to play Elizabeth I in a historical documentary.

No longer relevant: New, more directly relevant information can render other information useless.

> E.g. A person's sex may be relevant (statistically) to their body weight and relevant if body weight is important to the cooperation under consideration. However, as soon as you weigh them

[3] This is the optimistic view. Many given opportunities to show they have changed will do the opposite.

their sex provides no new information about body weight and sex becomes irrelevant.

Biased evidence: Evidence can be biased in many ways. The bias may lie in the details available about each person.

E.g. Many academic examinations require people to sit in an exam room and write by hand. Some people have faster handwriting and so have an advantage. They can write more and, often, that translates into more marks. However, almost all writing at work today is by typing into a computer. If speed of producing text is relevant at all it is typing speed that should be part of the test. Exams requiring typing in the exam room would be less biased. Alternatively students might be offered the choice of handwriting or typing.

Alternatively, bias may lie in the evidence about larger groups from which we learn how variables are connected.

E.g. We might have data on past employees that reveal links between their credentials when first employed and their subsequent performance in our organization. However, this only tells us about people we recruited, not the many people we rejected. That may eventually lead to bias in evaluating new candidates.

E.g. Suppose a dataset is used to price insurance policies based on the characteristics of customers applying online. But what if the data relate to past customers, many of whom did not apply online? It is likely that online customers and other customers are, on average, a bit different. Perhaps online customers are younger, more educated, or perhaps lazier?

Assessments can also be biased when evidence reflects actions taken in the past in response to assessments.

E.g. Suppose a bank assesses the probability of default for people asking for loans. It does this on the basis of its experience with people it has lent to in the past and the information they provided when they applied. Suppose also that, if a customer has a poor rating but still good enough to get a loan, the company takes extra steps to encourage the customer to pay promptly. If those extra steps work then those customers will pay promptly more often than they would have. That means that, if the company continues to base its assessments on experience in the same way, failing to take into account the extra effort to encourage prompt payment, their assessments will be biased.

Sometimes an assessment is automated by getting software to look at a set of past decisions taken by people and learn to replicate their approach. If those past assessments were unfairly biased then the automated assessments will probably be unfairly biased too.

Gathering evidence: The process of gathering information about a person is important. If we can easily gather more information about a person then it is unfair to judge them on a few weak clues and not bother with further information gathering. On the other hand, if gathering more information is difficult or impossible then it is fair to use what we have, even if it is not strongly predictive, provided we understand its limitations.

Being willing to find out more about people – to explore the possibility of cooperation – is fundamental to forming new, cooperative relationships. Without it many opportunities to cooperate beneficially are lost.

With that willingness to explore and develop possibilities, initial experiments with trading can develop into deeper relationships of more trusting cooperation.

However, it is still fair to put more effort into gathering information about people you are more likely to cooperate with. If, on the basis of initial information, a person looks a good candidate then information gathering will usually increase for that person.

Biased characterisation of a group: If we see a person before we hear or read about them then our initial information is based on their appearance. We usually know in an instant their sex, approximate age, physical attractiveness, race, ethnicity, and often their family wealth and fashion sense. It is normal, natural, and usually helpful to use this information. For example, knowing race can help with some medical diagnoses and is a cue to ethnicity, which helps with cultural sensitivity, serving digestible food, and some medical diagnoses.

A problem arises where the person's beliefs about a group (e.g. teenagers) are themselves unfairly biased (e.g. thinking they are all moody and difficult). For each characteristic of a person that may be relevant, the bias can be divided into two types:

- Bias about the average. For example, the person may think that teenagers are, on average, much wiser than the real average.
- Bias about the spread. For example, the person may think that teenagers are all very wise (spread is too narrow) or that their wisdom varies even more than in fact it does (spread is too wide).

Over-generalisation is the usual error. This typically means thinking that one group is better than another on the characteristic and that if you take a person at random from each group then you can be sure that the difference between the two individuals will be similar to the difference between the averages of their groups.

Failing to adjust: If the decision-maker has the opportunity to adjust their assessments in the face of new, relevant information then it is fair to do so. It's not unfair to make no adjustment if the new information arrives too late and commitments have already been made.

A clear case of unfair bias in assessment might involve a person unreasonably making no effort to gain further information and then refusing to even consider further information when it is offered to them and they are still in a position to adjust.

Fair reasons

Legitimate, fair reasons within an assessment of the suitability of a person for cooperation include the following:

- An objectively correct belief in the superior abilities or other qualities of one person over another (e.g. choosing to have surgery under a skilled surgeon; choosing the most competent player as a tennis doubles partner).
- An objectively reasonable inference of superiority based on incomplete information after diligent search for information.
- A family relationship, especially where there is a high degree of genetic similarity (e.g. parents, siblings, children). This makes them more likely to remain cooperative over time and so makes them more suited to cooperation. Family may also have an advantage in the final decision due to a wish to be kind to them, but this is outside the assessment.

- A romantic love relationship, again linked to the likelihood of continued cooperation.
- An established cooperative relationship (e.g. with friends, frequent collaborators, helpful neighbours).
- Greater knowledge of the individual, which helps to reduce our natural and reasonable wariness.
- Greater knowledge of people in a particular group, which also helps to reduce our natural and reasonable wariness.
- Having a strong cultural similarity, so that we are preferring to cooperate with people whose background is similar to ours and who are more likely to adhere to similar norms, know the same laws, and generally have an established pattern of cooperating with the society as a whole.

Some of these points need further explanation.

To be wary of people we do not know is reasonable and fair. Few people would happily agree to share a home with a person who is to be selected at random from the entire population of the Earth. Similarly, we would not be happy at the prospect of sharing a home with a person selected at random from the population of the UK, or our home town, or even the road where we live. Some people are nasty and even dangerous. Many people are unreliable or incompetent. So, naturally, we are wary.

It is not reasonable to expect people to react warmly to others they know nothing about. As we get to know people we rapidly revise our views of them, using every scrap of information. That information may make us more worried, or less.

Some of these legitimate, fair reasons within assessments of suitability can easily be mistaken for unfair bias. For example, preferring to work with people with whom we share a culture is reasonable if based on the advantages of familiarity and compatibility, reasonable if the culture we share is superior to that of other candidates, but unfair if based on a false belief that our culture is superior.

Can one culture be superior to another? Yes. Almost everyone thinks there are cultures that are better than others and this is demonstrated by the effort we put into debating changes to our culture, such as moral guidelines and laws. We are comparing two cultures – the one we have and a modification of it – and deciding which is best.

Two further issues can be important and may or may not be fair considerations, depending on factors not yet well understood. These are grey areas.

The first of these grey areas involves anticipated assessments by others. A person may say that someone is not suitable for a role because of anticipated assessments by others (typically customers or employees). For example, a 55 year old who wants to work in an office where all other workers are young females, an atheist who wants to be a political candidate for an election in a highly religious country, a person who wants to be a model but has severe scars from facial burns, or an actress who wants to play a sympathetic historical role but is not pretty.

Should society provide support in pursuit of wider societal goals?

> E.g. A shopkeeper owns a small shop in a town where most people are followers of a fundamentalist religious cult that believes gay people are inherently and deeply immoral. A candidate to work at the shop is

suitable but for being blatantly homosexual (through appearance and manner). The shopkeeper is not religious and in other circumstances would be quite happy to employ a gay man. Many today would say that if the shopkeeper considers this factor then her assessment would be unfairly biased. The problem is that if the shopkeeper hires the gay man then sales will fall and the business will fail. The cost of society pursuing its goal of fair assessments is borne almost entirely by the innocent shopkeeper.

Another type of problem arises where it is not clear if the anticipated assessments would be fair.

For example, is the pleasure of looking at physically attractive people a legitimate part of watching a movie, even when the plot does not specifically require them to be good looking? Consider Jaclyn Smith playing Florence Nightingale, Jenna Coleman as Queen Victoria, or Joseph Fiennes as William Shakespeare. These examples take things further because the historical figure was considerably less attractive looking than the person who played them in film.

What about the looks of competitors in sports? Professional sport is also an entertainment business. What about the person who serves you a drink at a bar? Or someone who sells you a vehicle?

When assessing attractiveness, does age matter? Is it legitimate to consider someone less attractive as they enter their 50s? I suspect that if many people were asked this in a large survey then the proportion thinking that age is a legitimate consideration would be higher for people in their 20s than for people in their 50s. But this does not settle anything.

Another problem can arise where the employer is concerned about some anticipated assessments but not others.

E.g. The two leading players in a successful TV action series have very different views on some hot political issues where public opinion is currently divided in half. The stars both post tweets making mild statements of their views. In both cases the reaction is a mixture of anger and support. The studio that employs the actors decides that one of the actors has posted 'disgusting hate' and starts to work out how to get rid of the actor. Internally, executives say they need to act to protect the popularity of the show. However, they ignore the tweets by the other actor.

The other grey area is where there is concern that the person being assessed will themselves make unfair assessments (and act on them) if given the role. How confident do we need to be that they will do so for this to be a significant issue?

E.g. A candidate shows they think they are part of an oppressed group and that, if they get the role, they will do things for others in their group to the detriment of others outside it. On its own this is not enough to determine that the person's actions would be unfair. We also need to know that their perception of the level of oppression is exaggerated and the actions would be excessive in relation to the actual level of oppression (if any).

E.g. A candidate is a member of a religious sect that teaches its followers, from childhood, that non-believers are immoral and inferior to believers in almost every way – little better than animals. Is this enough on its own to make the risk of unfair bias significant in an assessment? What if the person already has a track record of treating non-believers worse than believers?

A mathematical effect makes this situation quite likely if the person is a member of a

small demographic minority. Where there are incidents of conflict between people in two groups then individuals in the smaller group usually experience more conflict because the effects are concentrated in a smaller group. Consequently, they are more often victims than individuals in the larger group. However, they are also more likely to be perpetrators than they otherwise would have been.

Research challenges

Clearly, determining the extent of unfair bias within assessments of a person is made extremely difficult, perhaps at times impossible, by the difficulty of two tasks in particular:

- Deciding what is objectively correct, or objectively reasonable. For example, if two cultures are equally good (at least with respect to a particular type of cooperation) or one is better than the other.
- Separating the effects of legitimate, fair reasons for preferences from unfair biases. For example, separating the preference for a compatible, familiar cultural background from a false belief that one culture is better than another.

Nevertheless, if unfair bias in assessments of suitability for cooperation is to be demonstrated then these problems have to be solved.

As an example of the difficulty of deciding what is objectively reasonable, consider the Tudor period. At this time European culture included sophisticated architecture, food, clothing, music, and art. The Mona Lisa was painted during this period. Hampton Court Palace was constructed.

This was also the period during which Europeans discovered the Americas and a number of other parts of the world. There they found some non-white people living in ways that did not involve the technological and artistic sophistication developed in Europe. The limited but stark evidence at that time pointed towards those tribes being not only culturally behind Europe but probably also inherently inferior (i.e. genetically inferior to use modern language).

Examining bias in assessing non-white people at that time would have been difficult due to the very limited evidence available and the nature of that evidence.

Today the evidence base is very different. There has been vastly more contact between white and non-white populations. We are used to the idea that nearly all countries around the world have bustling cities with cars, aeroplanes, and the internet. Even supposedly 'remote' tribes shown in television documentaries are often seen wearing mass-produced T shirts. Most white Europeans have friends or colleagues who are not white, have the usual level of European education, and the usual cognitive abilities.

At the same time, theories about the impact of culture and education have grown in sophistication and credibility.

Today it is reasonable to think that there are no genetic differences in average mental abilities between races. Indeed, this is usually the assumed position in the absence of completely convincing evidence otherwise. Though there remain significant gaps in scientific knowledge in this area the evidence base is greatly improved and we can be fairly sure we are much closer to the truth.

Were Europeans in Tudor times evil racists and white supremacists? If someone held their views today we would probably say they were, but that is because the evidence base today is radically improved from Tudor times. Today we can do a much better job of determining the objective truth.

Research methods

A number of approaches can be taken to searching for unfair bias in assessments of people.

Before describing these approaches, here is one approach that is very often used but unreliable.

An unreliable method

One commonly used method that has fundamental problems is to consider one characteristic of the people cooperated with and compare that statistically to either the general population or those considered for cooperation and rejected.

For example, the inference might be something like this: 'Of the 20 people selected, 15 were supporters of Liverpool Football Club. That's 75% of those selected against a population average across the UK of only 2%. This clearly shows an unfair bias towards Liverpool Football Club supporters.'

Suspicious? Not if the recruiter is a sports betting company based in Liverpool. In that case, most candidates will be people with an interest in spectator sports, football in particular, who live in Liverpool and are typically supporters of the local team.

Far from being an occasional annoyance this kind of inference problem is almost inevitable in all cases. If people are selected for cooperation on highly relevant evidence and assessments with no bias then it will still be the case that on some of their irrelevant characteristics those selected are not typical of the general population.

This is partly due to coincidence. The more irrelevant factors you consider the more likely it is that you will find one on which there appears to be a large departure from the general population. Smaller sample sizes make extreme percentage departures more likely.

It also happens because characteristics that are not directly relevant in themselves will often correlate statistically with characteristics that are relevant. In the example, living near the employer and being interested in spectator sports are relevant to performance as an employee even though it is not helpful for them to be supporters of Liverpool Football Club specifically.

Finally, it can happen because the assessment is not biased and a group deserves its poor average ratings.

Reviews

The following activities for studying possible unfair bias in assessments of people do not involve statistical analysis. Instead they involve reviewing information about the assessors and approach used.

Usually, review is essential in order to establish that a bias is *unfair*. Statistical analysis is more useful for finding the bias.

Circumstances and behaviour

Establishing that an assessment is unfairly biased as opposed to merely biased can be difficult. However, useful evidence can come from looking at the circumstances around the assessment and the behaviour of those doing or controlling the assessment (e.g. writing interview questions, coding statistical analysis).

These facts may raise the suspicion of unfair bias but on their own will very rarely show unfair bias. If the risk is high then we should study the assessment methods and evidence used more carefully. If that shows bias then it is more likely that the bias is unfair rather than just the result of error despite an honest and diligent effort.

Vested interests are an important source of unfair bias. At stake may be monetary, material but not monetary, reputation, or relationships.

Another cause for concern would be if the time, effort, and competence applied to the assessment falls far short of what is reasonable.

> E.g. If a company gives important assessment tasks to someone who has no competence in the assessments, no time to do them properly, and makes little effort then there is a good chance that any bias found is unfair.

For many assessments there are already well-established good practices and known pitfalls with known corrections. If an assessment falls short of these good practices despite ample resources or fails to avoid mistakes that are well known then any bias resulting is probably unfair.

People doing the assessment might have irrational, perhaps emotional reasons for bias. Knowing their past experiences and assessments might reveal potential causes of unfair bias.

> E.g. A person who has had a series of bad experiences with red-haired people might have negative feelings towards redheads even when hair colour is irrelevant to an assessment.
>
> E.g. A devout follower of a religion that asserts that money lenders are evil may have irrationally negative views about bankers, perhaps even extending to the families of bankers.

Finally, an assessor with a track record of unfairly biased assessments is more likely to produce unfairly biased assessments in future. This is not someone who made a mistake once and has learned from it.

As mentioned earlier, this sort of review rarely *proves* unfair bias, even though it may raise the suspicion level. We need to establish that there was bias too, and that is not evident from vested interests, emotional issues, laziness, or a poor track record.

> E.g. It is not appropriate to attack an assessment simply because the assessor once tweeted something that might be interpreted as sexist, or had an aunt who was a Nazi, follows a religion, or regularly reads a particular newspaper. This is not a persistent track record of unfair behaviour and there must also be a biased assessment before there can be an unfairly biased assessment.

Evidence and reasoning

Reviews can also look at the evidence and reasoning behind an assessment.

Evidence may be pre-existing, such as previously recorded age, height, past examination results, and past criminal convictions. Or the evidence might be captured specially for the assessment, such as the results of a test administered to support the assessment.

Specially designed tests usually provide better evidence.

In either case, the evidence may be flawed in two possible ways:

- by simple mistakes
- through lack of *relevance*.

Mistakes might result in wrong numbers from faulty writing or calculations, missing data, or even duplicated data. Such mistakes are almost certainly more common than most people realise, so are well worth checking for if possible.

Relevance is an even more common area where problems can arise. Unfair bias would be indicated by:

- using evidence that plainly will be irrelevant
- not using obviously relevant evidence that was available or could easily have been generated.

It helps to distinguish between direct, causal relevance and mere statistical

relevance. To illustrate the difference, here is an example.

> E.g. Suppose that a town has a tiny table tennis club based in a large shed but run by a former international player with a rare talent for coaching young players. Within a few years several of the best players in the country are current or former members of the club. Soon most of the country's international team are alumni. The direct, causal reasons for their great play are the excellent coaching and intense practice. Living in the town, which most members do, is only statistically relevant to being an excellent table tennis player. Statistically, the town is far ahead of any other part of the country but of course there is nothing else special about the town itself. Residents who are not in the club are no better at table tennis than non-players anywhere else.

A direct, causal relevance means that anyone who has the attribute is affected by it to at least some extent. Merely statistical relevance does not have the same universal impact. There isn't a direct reason why it is relevant.

> E.g. Suppose a large number of people apply for a large number of computer programming jobs at a company over a period of years and some are successful. The main recruiting manager is an enthusiastic dog lover and programmer but dogs have nothing to do with the jobs and loving dogs is not statistically related to being a good programmer. Suppose also that the decisions on who to hire have all been made using the same algorithm and the same data fields. If those data fields include 'Dog Lover' then the suspicion is that this completely irrelevant variable is a cause of unfair bias. If the data fields do not include 'Dog Lover' but do include 'Pet Owner' then this is also irrelevant and just 'Dog Lover' in disguise because it will correlate very highly. This is unfair. In contrast, if it happened to be that pet owners really were more often good at programming then using pet ownership might well be fair, though using a more direct variable that explains why would be better.

A careful distinction is needed between:

- variables that are not relevant to prediction and do not help an objective assessment algorithm predict more effectively; and
- variables that we wish were not relevant to prediction but are.

A variable that is not relevant to prediction should be excluded from assessment but a variable that is relevant (directly or statistically) should be included, even if we wish it was not relevant. Why?

If a factor is a disadvantage for a person but not their fault (e.g. parental influence, a past illness) then it may have hindered their opportunities and progress in life. They may have greater potential than their achievements to date suggest, unless that disadvantage is taken into consideration.

Kleinberg et al (2018) explain the logic of this in detail and illustrate it with predictions using real-world data. They also explain that if the resulting assessment still does not provide the help that a particular group is thought to deserve then the best approach is still to use the best assessment possible through using all the data. The extra help is provided by using different criteria for different groups.

Corbett-Davies and Goel (2018) also argue for including variables that have predictive

value, even if they indicate 'protected' group status.

A variable that has statistical relevance only may be rendered useless by considering a second, more directly relevant variable.

> E.g. Suppose the initial idea is to look at sex as an indicator of physical strength, because on average men are stronger than women. Adding a test of physical strength to the assessment renders sex irrelevant.

We can discover the evidence and reasoning used in a number of ways. It is easiest where assessments are fully automated using software and it is easy to see what logic and evidence (data tables and variables) were used.

> E.g. If you want to be sure that an assessment did not respond unfairly to a person's physical attractiveness it is reassuring to find that the data used did not include any variable representing physical attractiveness, or a close proxy for it.
>
> If the software nevertheless selects an unusually high proportion of very good looking people then that will be because good looks are statistically associated with something relevant that the software was looking for. The software was not unfairly biased.
>
> E.g. A study by Cook et al (2018) of Uber drivers in the USA found that men were, on average, paid more than women each week and that the men were paid at a slightly higher rate per hour. However, the software did not use sex as a variable and the differences were due to men, on average, driving faster than women, doing more driving for Uber and acquiring more skill as a result, and driving in better paid locations.

If assessments are not automated we might be able to discover the reasons and evidence used by asking the assessor. In other situations there may be documents that help, such as instructions to assessors on what evidence to use and how to consider it, or records of the evidence used and the reasoning applied. These methods of finding the reasoning and evidence used are less reliable than reviewing software code.

Relevance is also a key consideration when looking at specially designed tests used to provide evidence for assessments. For example, it might be helpful to look at the person's physical fitness, their skills, or their honesty in particular situations using some kind of test.

Objective tests of a person will usually be a step towards greater accuracy and reduced bias. It is usually better than trying to estimate a person's abilities by reading about what they have done in the past or hearing them talk about their intentions for the future.

However, tests can be misleading and introduce bias of their own. If there is no good reason for the bias, or if no attempt is made to correct a problem pointed out, even though correction is not difficult, then the bias is unfair.

> E.g. A test of physical fitness might measure strength by finding out the heaviest weight a person can lift off the ground once. However, if the cooperation would involve lifting a lighter weight off the ground about 50 times in one day, then the test is likely to be biased towards people with a lot of strength but less endurance.

In general, the more accurately a test matches the performance required in the cooperative activity the less the bias. This is relevance again.

E.g. If the cooperation requires lifting a weight of 25kg off the ground around 50 times in the course of one day then a test that requires doing just that will be less biased than one with a different weight, a different number of repetitions, or a different time period.

A test that assesses a person's ability to improve can provide crucial information that helps reduce bias.

E.g. Two people might have identical strength now even though one has much greater potential for gaining strength due to their build and physiology. A test that assesses this potential will improve the decision. With such a test, high potential people who have not previously had the opportunity to train hard may emerge as better candidates for cooperation.

Some advantages enjoyed early in life (e.g. being one of the oldest in your year at school) can give an edge that lasts for years because high performance is rewarded by greater opportunities to develop further.

Tests of improvability (e.g. response to physical training, learning speed) can be valuable in identifying people who are long term good choices but so far have not had the opportunity to develop. This is important in realising the productive potential of people.

Failure to test improvability may be an instance of unfair bias.

Reliable tests of honesty are extremely difficult to design because, of course, liars can cheat at tests. However, two exceptions to this come to mind.

Some surveys include questions designed to identify people with a strong desire to give what they think are socially acceptable answers. A question might be 'Have you ever withheld information simply because it was embarrassing to you?' The answer 'No' is almost certainly a lie. We all have. Several such answers to similar questions indicate someone whose other answers cannot be trusted.

Features to look for
Assessments that are forecasts, because they are easier to compare with reality.
Explicit methods, automated if practical.
Trying to produce assessments (predictions) that are as accurate as possible, even if equity considerations require decisions that are not consistent with the assessments.
Using all data that can help make a better prediction, including sensitive characteristics, if and only if they improve predictions.
Trying to render sensitive characteristics irrelevant by using more directly relevant variables.
In particular, devising realistic tests of people that replace inferences from demographic or biographic data.
Considering improvability as well as current performance level.
Testing the accuracy of predictions against reality (preferably using data not used to build the assessment model).

Table 2. Good features that contribute to fair assessments of people.

Some corruption is by people who believe that family and friendship bonds are more important than society's rules. In this way of thinking, if your cousin wants a job and you can give it (despite not owning the business) then it is your moral duty to do so even if other candidates would be better for the business. People with this type of morality sometimes express it openly because in their view there is nothing wrong in it.

Table 2 summarises some good features to look for in assessments of people. They tend to promote fair assessments.

Group profiles used

Many assessments are influenced by pre-existing knowledge of people within groups (i.e. having a particular characteristic such as hair colour, age, sex, etc).

A broad strategy for detecting unfair bias in those assessments, within which there are many possible methods, is to focus on perceptions of group characteristics to see if they:

- agree with reality;
- reflect the evidence that a person reasonably should have considered; and
- move towards reality when credible factual information is provided about the group (showing that people are updating their views).

When asking how people see groups, be careful to allow them to give rational, fair answers without ambiguity.

In practice many surveys and other tests designed to measure something like unfair bias do not meet these simple requirements. In the following list of common faults a continuing example is used to make the points of principle and this is supplemented by examples from well-known instruments[4] used by academics and other surveys.

As an illustrative example, imagine we are trying to establish if a person has an unfairly biased view of the intelligence of serious chess players based in the UK. Perhaps they over-estimate their intelligence, or perhaps they under-estimate it. Perhaps their view is inconsistent with whatever evidence is available. Perhaps, when confronted by detailed numbers on the intelligence of the UK's top chess players, the person sticks with a view that is very much inconsistent with the facts.

Here are some potential problems with instruments used to measure unfair bias:

No comparison with reality: To establish that the person has a view that is biased (for any reason) we need reliable information about the group property in question. In our illustration, that would be the distribution of intelligence within the UK's top chess players. If we don't know this then we cannot identify if a person's views are biased.

A common failing is to make no effort to assess reality or compare it to the person's potentially-biased views.

> E.g. The Pro-Black Scale (Katz & Mass, 1988) asks for the extent of agreement or disagreement with various statements including 'Most blacks are no longer discriminated against.' The proportion of black people discriminated against (unfairly, presumably) is something that depends on when and where you ask the question. The scale was first published in 1988 and considerable changes have taken place in many countries since then. In some countries today the statement is almost certainly true so agreeing with it cannot be a sign of racism (the aim of the scale).
>
> E.g. The Ambivalence Towards Men Inventory (Glick and Fiske, 1999) asks for extent of agreement with the statement: 'When in positions of power, men sexually harass women.' Although it is hard to believe that literally all men in positions of power sexually harass women, the research behind the scale includes no attempt to even estimate the proportion of men

[4] In this context 'instrument' usually means a questionnaire, but sometimes more elaborate methods are used.

at any level of power who sexually harass women. Such research would be difficult.

E.g. In contrast to these poor examples, McCauley and Stitt (1978) in their third study asked subjects seven pairs of questions that asked for well-defined estimates about defined populations. The first pair of questions was 'What percent of adult Americans have completed high school?' and 'What percent of adult American blacks have completed high school?' The other pairs asked similar factual questions about other percentages for the two populations. The answers were compared to recent official government statistics. Although estimates were similar to the real rates the tendency was to underestimate the difference between black adult Americans and adult Americans generally.

In some cases the developers of a scale may have felt that they knew the truth already. However, it is important to note that their procedure does not involve an explicit comparison with the truth or any attempt to establish the truth. As time passes an instrument that initially seemed to indicate unfair bias may become invalid because reality has changed. The scale should identify this.

E.g. McConahay's Modern Racism Scale (1986 – so not modern now) was based on the idea that modern racists (assumed to be white people in the USA) thought that racism was no longer a problem and that too much was being done for black people. In principle, as societies continue to change, there may well be a time when racism in a particular country really is no longer a problem and that too much is being done for black people. The scale and its associated procedure would still report that white people are racist and that is a fundamental problem, in principle, with its design.

Only allowed to err in one direction: A test of unfair bias should be able to detect bias in either direction. In our illustration using chess players, that would be over-estimating intelligence as well as under-estimating it.

E.g. With the Johnson and Lecci Scale for anti-white attitudes (Johnson and Lecci, 2003) it is mathematically impossible for a respondent to register a positive attitude towards white people on any individual item, still less overall. The most they can do is show no negative attitudes. The same one-sided design is present in Brigham's Attitude Towards Blacks scale (1993), McConahay's Old Fashioned Racism and Modern Racism scales, Katz and Mass's Pro-Black Scale and Anti-Black Scales (1988), and Glick and Fiske's Ambivalence Towards Men Inventory (1999).

E.g. In contrast, Ashton and Esses (1999) asked subjects to estimate the average academic grades of Canadian high school students in nine ethnic groups. These estimates were then compared with reality. The estimates could have been too high as well as too low.

Nothing comparable with reality: The person's responses should be comparable with reality so that a comparison with reality can be made. In our illustration, that might mean the person being asked to estimate the average IQ score of top UK chess players, or give more details about the distribution of IQ scores for that group.

An often-used method is to ask people how much they agree or disagree with a statement e.g. 'Professional chess players are highly intelligent.' The agreement is

expressed by choosing a position on a scale ranging from 'agree strongly' to 'disagree strongly'. This cannot be compared to reality.

This style of question is called a Likert item. Likert items are extremely common in psychology but the scores that come from them do not quantify anything that is objective or comparable with reality.

> E.g. Likert items are used in Brigham's Attitude Towards Blacks scale (1993), McConahay's Old Fashioned Racism and Modern Racism scales, Katz and Mass's Pro-Black Scale and Anti-Black Scales (1988), the Johnson and Lecci Scale for anti-white attitudes (Johnson and Lecci, 2003), and Glick and Fiske's Ambivalence Towards Men Inventory (1999).

Confusing uncertainty, frequency, and emotion: Another consequence of Likert items is that people are unsure how to answer in response to statements whose truth is uncertain for them. For example, if asked for their degree of agreement with 'Most professional chess players have an IQ of more than 120.' they may be unsure of the truth. The statement is either true or false, not true to a degree. How are they supposed to answer? Does 'agree strongly' mean that they are very confident that the statement is true, or does it mean that they think it is true and they feel very strongly about it? Uncertainty and emotion have become confused.

Another ambiguity arises when it is obvious that the statement is not true for all members of a group so perhaps the question is asking how frequently the statement is true.

> E.g. Glick and Fiske's (1999) Ambivalence Towards Men Inventory asks for agreement/disagreement to the statement: 'Men pay lip service to equality, but can't handle it.' Since this is unlikely to be true for all men an alternative interpretation is that the questionnaire is really asking what proportion of men the respondent thinks cannot handle equality.

Pushing people into generalisations: Typically, characteristics vary within a group and questions should allow people to show that they understand this. In our illustration, serious chess players will have a wide range of intelligence levels, so it is wrong to ask people to consider generalisations that apply equally to all players such as 'chess players are smart'.

> E.g. Pettigrew and Meertens's Blatant and Subtle Prejudice scale (1995) asks for degree of agreement or disagreement with many statements, one of which is 'West Indians living here teach their children values and skills different from those required to be successful in Britain.' Giving this rating requires a generalisation across all West Indians living in Britain. As with any large, demographically defined group, some West Indians living in Britain do indeed teach their children values and skills different from those required to be successful in Britain.
>
> E.g. Johnson and Lecci's anti-white attitudes scale (2003) includes the item 'I believe that the success of a White person is due to their color.' This is asking for a generalisation across all white people who have achieved any kind of success.

Presenting unclear propositions or asking unclear questions: Propositions about groups need to be clearly stated or their factual accuracy cannot be assessed and it is unclear what the respondent truly thinks.

A typical mistake is failing to identify the group involved through using a general term like 'chess players' or 'black people'.

Does this mean all chess players in the world? All chess players in the Western world? The UK? All chess players the person knows about? If a person is in frequent daily contact with unusually dull-minded chess players then the factually accurate answer for chess players they know will be different from the factually accurate answer for all the world's chess players.

> E.g. The American National Electoral Survey Pilot Study (2016) asks for an overall feeling towards various groups of people. As an example, one of the items says 'How would you rate scientists?'

Another mistake is to use quantitatively vague terms. For example, to say that chess players are 'intelligent' is quantitatively vague. How intelligent is 'intelligent'?

> E.g. In the American National Electoral Survey (2016), a question about groups asks 'How well does the word "lazy" describe most members of each group?' How 'lazy' is 'lazy'?

Yet another all-too-common mistake is to use items that rely on the person understanding words that the researcher thinks are clear but do not have a clear and generally agreed meaning. Words like 'stereotype', 'bigotry', 'prejudice', and 'discrimination' fall into this category.

- The Old-Fashioned Racism Scale asks for the extent of agreement or disagreement to a number of statements including 'Generally speaking, I favor full racial integration.' What exactly does 'racial integration' mean to most people?

Confusing lack of familiarity with unfair bias: Not knowing much about a group of people is not in itself evidence of unfair bias towards them. Perhaps you grew up in one country and so know most about the people of that country and much less about people of other countries. Some instruments take lack of familiarity as evidence of unfair bias.

> E.g. The Attitude Toward Blacks Scale (Brigham, 1993) has 20 items and people have to say how much they agree or disagree with them. One of the items is 'I think black people look more similar to each other than white people do.' People develop the ability to distinguish between visually similar objects (people, flowers, seashells, etc) through exposure to them and effort to learn to identify them. A person who grows up with people of a particular race will be better at distinguishing people of that race and see individuals as more different from each other. Lack of this ability due to relative unfamiliarity is not a reliable sign of unfair bias.

Many of the items on the long-established instruments for assessing bias are somewhat political statements that most people will approach warily. To be polite and reduce their risk, respondents give the answers least likely to get a disapproving reaction (even if it will never be seen by the respondent). In response to this the researchers in this field turned to trying to measure the extent to which respondents were motivated to give socially acceptable answers. They also searched for methods that respondents could not fake. The Implicit Association Test was developed for this purpose but has significant problems.

Confounding unfair bias with other preferences: The complex procedure of the Implicit Association Test eventually produces a difference in average decision reaction time when people are asked to press one of two buttons when presented with a word. The Implicit Association Test gives people a task that involves them

having to remember which two attributes are associated with each of two buttons, and remember new associations as they change during the test. It is very likely that most people try to link the two attributes in some way to help them with this task. The test shows the relative ease of *making* that association, not that people have already made that association.

Many things can make that association easier that have nothing to do with animosity or a desire to do harm.

> E.g. In one study by Greenwald et al (1998) Korean people were asked to press one button if the word on the screen was a Korean name and another button if it was a Japanese name, but if the word was not a name then they had to press one button if the word was associated with something pleasant and the other if unpleasant. On average the Koreans were slightly slower to make the decisions if pleasant and Japanese were using the same button. What does this mean? The researchers thought it meant that the Koreans did not like Japanese people but it could also have been that the Koreans did not like Japanese names – in the sense that they would not give those names to their own children and did not have many friends with Japanese names.
>
> E.g. The same paper reports another experiment where white Americans were asked to press buttons to decide if the names where typical white or typical black names. The names had been chosen by frequency of occurrence in a local directory and for being characteristically black or white. What they did not do was equate the apparent socio-economic status of the names. Several examples given are names that probably would not have been chosen by aspiring, college-educated black parents. This means that the socio-economic status of the names is confounded with race. Since only white students were used as subjects in this study it is hard to interpret the results.

Other explanations that have been suggested are that people more easily associate particular races with negativity because they see them as having been badly treated or being badly off, not because they see them as bad people. Alternatively, they perhaps are aware of often-mentioned cultural stereotypes or views without endorsing them. Or perhaps they associate their own race with positive concepts more easily because their favourite people (often family) are of their own race.

The Implicit Association Test has become the centre of some academic controversy lasting many years. However, these methodological issues have not been addressed and the procedure is still used and considered to be a valid measure of bias by its supporters. Consequently, it remains likely that some people are incorrectly identified as 'racist' by this test.

The assessments

Last but not least, the assessments themselves are sometimes so implausible that bias is obvious and probably unfair. To counter the obvious risk of misjudging what is 'implausible' this should only be done by recognizing particular types of claim that are made often and are nearly always false.

In particular, strongly worded generalisations about all members of a large group of people are almost always wrong.

> E.g. It would be wrong to say that 'people over 60 do not understand the internet.' The claim is too strong and

applies to all members of a huge group. There are bound to be exceptions. In fact this age group includes Sir Tim Berners-Lee, the inventor of the World Wide Web, and Bill Gates. A UK survey by the ONS (2019) found that 83% of people aged between 64 and 75 had used the internet in the previous 3 months, along with 47% of those 75 or over.

Generalisations about the behaviours and beliefs of people defined by their race, ethnicity, sex, nationality, home town, or anything else not specifically defined by the characteristic being assessed are likely to be wrong, provided the group is not very small.

Very few generalisations about all members of a group can be made. Sometimes a criterion for group membership seems like it would guarantee a related generalisation but even here it is unsafe to generalise with any large group.

E.g. According to a Pew Research Centre survey in the USA in 2014, 2% of Catholics did not believe in god and, overall, 36% were not certain there is a god.

E.g. MENSA 'The high IQ society' has one condition for membership, which is that you demonstrate that your IQ is in the top 2%. The problem is that people get better at IQ tests by practice and if you practise enough on the items that MENSA uses then you have a good chance of getting in, even if your initial scores (the ones psychologists typically use for IQ measurement) were not good enough. How many current members really are in the top 2%? Probably most but almost certainly not all.

The fact that these over-generalisations are so unlikely to be true, and often can be found to be false with ease, makes them unfairly biased. This is because, at the very least, they are negligent. If they appear alongside several other over-generalisations, slanted presentations of information, and incorrect statistical inferences then you can be sure that unfair bias is at work.

This is the easiest, quickest, and most reliable way to identify unfair bias in many cases. It is useful for many of the claims people make in politics but even academic papers and books sometimes include implausible generalisations. Leitch (2019) discusses in detail the problem of accidental over-generalisation through careless language and how to avoid it.

Statistical methods

Vary the variables

If the assessment is done through a statistical method with training data and all this is available to the person checking for bias then a simple approach is available. This is to examine the effect of excluding variables from the model and perhaps also adding variables (if there are data not already used). If either of these changes gives assessments that are better predictions of the outcomes then the new assessment is preferable and the old assessment was probably biased, though not necessarily unfairly.

E.g. In a general linear model regression with prediction performance evaluated using information gain it might be that adding a variable previously left out improves prediction performance. If this variable is a sensitive variable, remember that it is better to provide equity by choosing actions that are not directly based on the assessment, than by distorting the assessment.

It is less likely that taking out a variable will improve assessment performance. However, this might happen if the model

fitting criterion used did not match the assessment evaluation criterion, or if someone tweaked the model manually so that it was not exactly a model built from training data. It can also happen through reducing over-fitting or if the criterion used to judge performance penalises models with larger numbers of variables.

A deliberately tweaked model would be evidence of unfair bias.

Compare assessments with the truth

Assessments that are predictions can sometimes be compared with actual outcomes, once these are known, to detect bias. In other situations it may be possible to compare assessments with unbiased assessments made independently. (Though these are unlikely to be as reliable as comparison with actual outcomes.)

Some excellent mathematical formulae have been developed for comparing predictions with reality, including some for probabilistic forecasts. These include calibration and proper scoring rules such as information gain.

These metrics do not usually separate forecast error due to random, non-systematic bias from bias. Nor do they separate out *unfair* bias.

However, most forecast error is systematic and so most is bias (though not necessarily unfair). This is particularly true for automated assessments. When we look at forecast errors (i.e. the differences between forecasts and reality) it can seem that the errors are randomly distributed. Indeed, a typical requirement for a well-fitting statistical model is that the errors appear to be randomly distributed and evenly so.

But, in reality, the reason the forecasts are not perfect is typically because more information and understanding is needed to reach perfection. It may be that, at some quantum level perhaps, there is randomness that never, ever can be penetrated, but in practice we are almost always stopped from progress long before reaching any such randomness barrier. More often we just don't understand why our forecasts are wrong.

The main modelling imperfections all give systematic errors that constitute bias:

- **Missing variables**: The most common reason for poor predictions is lack of relevant information. Missing a variable means there may be people who would have got a more favourable assessment if only that variable had been considered.
- **Irrelevant variables**: A variable that is included in a model but makes the predictions worse is not as common but if present it will mean that some people are being penalised because of a characteristic that has nothing to do with performance.
- **Wrong model formula**: If a variable is related to the prediction using the wrong type of relationship (e.g. linear when exponential would be better) or variables are combined in the wrong way (e.g. additive when multiplicative would be better) then there will be systematic forecasting errors. People whose characteristics are at different points on each scale will be systematically advantaged or disadvantaged.

Therefore, these well-established measures of forecasting skill can be used to gauge bias.

To establish *unfair* bias it will usually be necessary to show bias *and* use knowledge of the circumstances to show that it was unfair. For example, failure to use a variable that improves predictions might be a basic error that could and should easily have been avoided. This

would indicate negligence or deliberately poor forecasting. Alternatively, the omission could be the result of that data not being available without a huge effort, not justified by the circumstances.

An assessment is free of unfair bias if assessments of each person are as accurate as can reasonably be achieved. That is, as accurate as our data, skills, tools, and knowledge allow. This does not necessarily mean that all subsets of the assessments are equally accurate.

In particular, if error rates are compared across two groups of people who have been assessed and one group is larger than the other, then errors will typically be smaller with the larger group. Conversely, less experience to draw on means higher prediction errors. This is an unavoidable statistical effect and not the result of unfair bias.

It would not be fair to use the same number of data for each group if this means not using some data. This would be the fundamental error of making things worse for some people in order to achieve equality – bringing everyone down to the same level. It is appropriate to use all the data available If that means more accurate forecasts for larger groups then that is sensible, because more people would be affected by forecasting errors in the larger groups.

If one group is genuinely different from the other this may also affect the pattern of prediction errors in a systematic way that is not the result of unfair bias.

With real data sets it is rarely possible to satisfy the demands of all the criteria that have been suggested even when only one characteristic for grouping is considered. If the goal is to satisfy them for all sensitive characteristics the chances of success are even lower.

Some of the 'fairness' metrics[5] that have been suggested require the average assessment given to people in each group to be equal, even if this is unrealistic. This is the fundamental mistake of deliberately seeing the world in a distorted way.

In summary, applying most of the metrics designed to identify bias will not usually identify *unfair* bias and it is only in particularly simple, clear situations that an inference of unfair bias can be drawn. It will usually be more fruitful to look for sensible design features of the assessment approach. If they are not present then unfairness is more likely.

Check for changes with new information

One of the signs of unfairly biased assessments is a failure to change views in the face of strong evidence. This is a potentially powerful method that probably could be used more often.

The simplest, three step procedure would involve (1) a pre-test of a person's views, (2) the presentation of real, credible, evidence that should lead to a revision of the person's views if they are biased, and finally (3) a post-test of the person's views.

This may be an easier procedure than others because it is not necessary to know so precisely what really predicts performance. We only need to know that:

- the person's assessment was inaccurate and in what direction; and
- the evidence should have led to a revision in the correct direction.

Model assessments

Another research approach is to:

- perform a multiple regression on a set of people and their assessments to

[5] In the literature on fairness metrics no distinction is made between fair bias and unfair bias. All bias is classed as unfair even if it is despite diligent effort to achieve accuracy.

build a model that tries to predict the assessments made for each person;

- perform a multiple regression on a set of people and their actual performance that tries to predict reality for each person; and
- compare the two models.

If a variable predicts the assessments but not reality then it could be that the assessments are wrong because of using a variable that is in fact irrelevant. The nature of that variable might suggest unfair bias.

This technique again requires the assessments to be predictions and it requires knowledge of what actually happened later in order to build regression models. The population used for each regression perhaps does not need to be the same but it would be ideal if it was.

Both regressions need to try out all potentially relevant variables that are available, including sensitive variables.

The regression models need to be of the same type. They also need to be of a type that includes parameters that represent the importance and use of each variable in making assessments. This is true, for example, with multiple linear regression but not usually with neural networks.

Unfortunately, establishing the parameter values with any precision and confidence requires a lot of data. There is often a problem when relationships are not linear. Also, regressions (inevitably) struggle to separate the importance of variables that tend to move together (known as 'multicollinearity'). You might see some weight attached to a variable that should be irrelevant, but this might be just confusion with a relevant variable.

Model hypothetical assessments

Another way to study a decision-maker's approach is to experiment with hypothetical decisions carefully constructed to uncover the way the decisions are made. Done properly this can separate reasons much more cleanly than regression using real-life assessments, but there are challenges.

Again, the variables used should distinguish carefully between fair and unfair reasons. It is possible to separate variables that usually move together, getting around the 'multicollinearity' problem that usually afflicts multiple regression.

One challenge is to ensure that information intended to be irrelevant is truly irrelevant and does not provide statistically relevant decision cues.

> E.g. A very common method of trying to identify unfair bias in hiring decisions is to use mocked up job application forms where all details are the same except for one, such as the race or sex of the applicant. Great care is needed to avoid providing additional information unintentionally. For example, showing a photograph of the candidate to communicate their race may also communicate such things as their physical attractiveness, strength, dress sense, and socio-economic status.

Again, this requires decision-makers to behave normally and to participate honestly, even though they probably know that the objective is to uncover unfair bias.

Consider uncertainty

Assessments of people are usually uncertain to some extent. It may be easy enough to measure how tall a person is but how about predicting their honesty in a new situation, or how successful they will be on a study course, or whether they will thrive as a lumberjack?

One obvious effect is that predictions based on lots of data are usually more accurate than predictions based on few data, other things being equal.

A less obvious but still common and rational response to uncertainty is called 'regression to the mean.' When predicting uncertain future performance of individuals within a group it is rational for the assessments of each individual to lie between the past performance of the individual and the average of the whole group. In other words, when we are unsure we shift a little towards what is average. This usually improves the average performance of predictions in the sense that outcomes more often match our expectations.

However, when you look at the predictions awarded to each outcome a different picture emerges. Now it seems that the predictions have a tendency to be conservative. That is, predictions for high actuals tend to be too low while predictions for low actuals tend to be too high.

When studying possible bias in assessments of people it is important to understand this effect and not confuse it with unfair bias.

> E.g. Wyness (2016) reports analyses of a huge database of university applicants in the UK, comparing their actual A level points with those predicted by their schools in advance. Consistent with regression to the mean, when average predictions for each actual points total were calculated the tendency was to over-predict results for students who went on to score badly in the exams, but under-predict for students who later did well. It also may explain why grammar schools less often under-predicted for students who went on to do very well in the exams – since the mean for grammar schools is higher than for other schools. There was also a slight tendency for students in locations where going to university is less common to be under-predicted, consistent with a lower average attainment in those locations.
>
> Unfortunately, Wyness does not consider the role of uncertainty in predictions or in the A level results themselves, and does not take into account the regression to the mean effect. When this effect is considered, exceptions to this pattern stand out that should have been highlights of the study but were all but ignored.

Compare assessors

If there are many people making assessments then it may be possible to compare assessments between people and identify individuals whose views are biased. Clues to why they are biased might even point to unfair bias.

If one demographically defined group has, on average, a different assessment than another, of the same people, then probably there is some bias and some of it may be unfair.

> E.g. Rudman and Goodwin (2004) used a slightly modified Implicit Attitude Test to assess attitudes of male and female psychology students at a university in the USA to men and women generally. In all four experiments the IAT scores suggested that women preferred women much more than men preferred men. If the IAT scores mean what they are supposed to mean then this indicates bias linked to sex, so probably unfair. However, since the measures cannot be compared to reality there is no way to know from these studies which sexes were (on average) biased.

E.g. Johnson et al (2019) used a more complicated regression approach to investigate fatal shootings by police officers in 2015 across the USA. This amounted to comparing officer behaviour by race while statistically controlling for some other factors. They looked at whether the racial mix of people shot by officers depended on the race of the officer. If, for example, white officers sometimes hate black civilians in particular then one would expect that the racial mix of people shot by white officers would swing towards more black people being shot dead. In the data set this was not the case. The racial mix of people shot dead was not changed by the race of the officer.

This second example is about actions taken, not assessments of people directly, but it illustrates a technique that can be used.

Incidents of bias

Another approach to studying unfair bias statistically is to try to count specific incidents of unfair bias. An example of an unfair bias incident is a 'hate crime' or crime that is 'aggravated' by a 'hate' element.

Not all incidents of unfair bias involve an unfairly biased assessment.

Research challenges

With this approach there is often a problem with biased measurement. This can change the appearance of trends and the different levels of unfair bias experienced by different groups. Reasons for mis-measurement include the following:

- Willingness to report incidents varies over time and between groups.
- The criteria for counting something as a bias incident can change over time or be different between different groups.
- It can be hard to know for sure if unfair bias was the reason for an incident. Where alternative interpretations are possible the number of people who assume unfair bias may change over time and differ between groups.

There can be ambiguity as to the true nature of an incident whether or not there is an explicit indication of bias. An explicit indication would be something like a statement of the reasons for a decision that unambiguously gives a reason that is irrelevant or states an assessed level that is objectively wrong, along with irrelevant reasons for the assessment.

With no explicit indication of unfair bias it is particularly hard to determine if unfair bias was a factor without access to the mind of the decision-maker. The person who may have been biased is in a particularly good position to know if unfair bias was a factor because they have some access to their own thoughts and so know their motives and considerations. The person who thinks they may have suffered from unfair bias does not have that special access.

People who have reached an unfair assessment will often avoid giving clues that would reveal what they have done.

E.g. If you are rejected from a job you seemed well qualified for the recruiter may provide what they call 'feedback'. But would they tell you if they rejected you because of a graphologist's analysis of your handwriting? Even people who think graphology works will usually know it has a (well deserved) reputation as useless pseudoscience.

With an explicit indication of unfair bias there can still be doubt in some cases. When people are frustrated and angry they reach for insults and what comes to mind may suggest bias that is not present. For example, suppose a slightly overweight woman makes a foolish and careless mistake, again, causing a co-worker to say 'You stupid fat cow.' These insulting, hurtful words may be just insults chosen in the heat of the moment. But they could also be taken as explicit indications of ableism ('stupid'), of fat shaming ('fat'), and sexism ('cow').

> E.g. Metcalf and Rolfe (2010) reported 19 cases based on interviews with people who believed they had suffered caste discrimination in the UK. The set of cases is only a subset of the cases they initially interviewed and one of the reasons for excluding some cases was uncertainty over whether there had been caste discrimination. Even with the cases reported it is often unclear whether the person thought to have discriminated did so because of caste prejudice or because they had some other reason but expressed their anger, disapproval, and so on in caste terms.
>
> E.g. Former Islamist Maajid Nawaz was once punched by a school friend who said it was because he was Pakistani. At the time Nawaz took this to be the true reason but years later learned that the real reason for the punch was jealousy over a girl (Nawaz, 2012).

Research methods

Logging incident reports

One approach with major problems is to ask people to report incidents and log details of the incidents they report. This is often the by-product of enforcement.

> E.g. In the UK crimes reported to the police are recorded regardless of whether any action is later taken. The statistics on 'police recorded crime' are available in detail and for a huge number of incidents across the whole country.
>
> But, despite the huge amount of data collected, the problems of inconsistent reporting and classification mean that it is impossible to say with certainty whether a particular type of crime is increasing or not, or if one part of the country experiences more of a crime than another. All you know is what has been reported and recorded.
>
> This applies in particular to what is called 'hate crime' and crimes aggravated by 'hate' of some kind. These are particularly susceptible to different levels of reporting of perceived incidents, for all the reasons already discussed above.

The same will apply to logs of most other forms of suspected unfair bias incident.

Survey data

A much better way to find out about the true prevalence of unfair bias incidents is to use random sampling and a survey. Respondents are selected at random and asked if they have experienced any of a list of precisely defined types of incident of unfair bias. Indeed, even if they cannot be sure that unfair bias was involved they can still report if they experienced an incident with particular, defined features.

Although this requires extra effort and does not involve the whole population, more reliable inferences can be drawn about trends and differences between groups, provided a very high response rate can be achieved (i.e. almost everyone responds).

> E.g. Crime statistics reported by the UK's Office for National Statistics are usually based on two sources: (1) police recorded crime, and (2) a

regular crime survey known as the CSEW (Crime Survey for England and Wales). Hate crime recorded by the police has been rising for some years but the crime survey has not shown the same increase. This is interpreted by the ONS as evidence that hate crime is not increasing, but reporting of hate crimes is increasing.

A survey will be more effective if care is taken over the wording of questions.

- Incident types should be defined in objective, practical terms, not abstractly and not in terms of perceived 'offence'.
- Incidents should be divided into different levels of severity to avoid a broad category being made to look worrying by a large number of very mild incidents.
- The focus should be on severe incidents, where memories and official records are likely to be more reliable.

Getting most people to respond to the survey is important. If there is a systematic difference between people who respond and people who do not then percentage incident rates from responders will not be representative of the whole population.

Differing outcomes

Another situation where people often try to identify unfair bias is where two groups (usually identified demographically) experience different outcomes on average. For example, where one group gets paid more than the other, on average.

This analysis is much more complex than studying decisions to cooperate, where the focus is on a specific type of decision. Now we are talking about a whole life, including decisions taken by the alleged victims, opportunities around them that are influenced by much more than unfair bias, and perhaps unfair bias too.

To take just one example of the complexities, imagine we are trying to see if differences in wealth between people of different ages are the result of unfair bias. An obvious reason why people of different ages have different wealth is that people usually earn money during their working lives and many save some of that money for their retirement or convert it into long lasting assets, mainly their home. So, in many cases, people get wealthier as they get older up to a point near retirement where they start to get less wealthy again. In theory this on its own could create very large differences in wealth without any unfair bias being involved.

The differing outcomes also reflect the behaviour of many more people. Instead of just some decision-makers and some people being assessed we have many more people whose behaviour is important. All these could be asked to behave differently if they are sometimes unfairly biased:

- The people whose outcomes are being compared.
- Their families, friends, peers, and advisors who might influence their behaviour.
- News media and other public opinion influencers who might influence those being compared.
- Decision makers who control or at least influence opportunities for the people being compared.

The influences on these people may include assessments of unfair bias, and these might be under- or over-estimates.

Unfair bias

The unfair, biased reasons for such differences include:

- Differences in opportunity imposed by one person or group on another for no good reason.
- Advice given to members of one group, for no good reason, that puts them at a disadvantage if they respond to that advice. This might be subtle, as with implied low expectations.
- A person being discouraged from taking a good opportunity by incorrectly thinking that the opportunity will be spoiled by others being unfairly biased.

All these might be the result of unfairly biased assessments of people.

Fair reasons

Reasons that do not amount to unfair bias are as follows:

- A difference in measurement that creates the false impression of a difference in outcomes.
- Choices freely made by group members, without unfairly biased influences.
- Differences in opportunities resulting from fair decisions made by others.
- Differences in performance by group members, for a given set of opportunities.
- Differences in recognition and rewards resulting from fair assessments and decisions made by others.

Behind some of these are some important influences:

Biology: Our genetically inherited potential is important to many of our abilities while being able to give birth to a child makes a huge difference to your life choices. Genetic potential affects how productive you are and how quickly you can adapt. It also affects choices you make because most people prefer to do things where they have a natural advantage.

Culture/technology: The abilities we acquire from schooling, learning at home, learning from friends, and other sources can be a great help to our progress and, again, influence what we choose to do.

Wealth: The wealth our parents have and pass on to us can have a profound effect on our opportunities.

Information: The more we know about what is going on and how the world works the better our chances of making wise decisions, provided we are not overwhelmed by information. Understanding, at a young age, how the adult world of employment works and how different qualifications can affect future opportunities and income is crucial. If all you have as a guide is what your friends are doing, and they are all choosing to study fine art, then you can expect to be poor until you change direction.

Research challenges

Analysis of the causes of differences such as differences in pay, occupations, and longevity is often made more difficult by unreasonable pressure to make assumptions for which there is little or no evidence. For example:

- To pretend that there is no difference when there is.
- To pretend there are no inherited, genetic differences when it is not known if there are differences or not.
- To pretend that the entire difference is the result of unfair bias, or assume it until someone else can prove it is not.

The idea that genetic differences are not involved may be well-intentioned but can cause problems and unhappiness. It may be driven by a desire to be encouraging – to get people to make more effort on their own behalf in the belief that they can do

anything if only they try hard enough. But when real progress is slow and others seem to make much faster progress, what explanations are left? Is the relative struggler lazy? Or is it that someone is unfairly making things hard for them? Ideally, we should recognize the existence of genetic advantages to precisely the extent that they really exist.

Unfortunately, we currently cannot directly assess a person's genetic make-up and their potential. We can only study their actualised abilities, and those of their parents, and try to allow for differences in experiences in some way.

Research methods

An unreliable method

A common but unreliable type of inference points out that two groups of people differ on one outcome. This can be done in a variety of ways but the key point is that only one characteristic – usually a group membership – is considered. These are the three main forms with some hypothetical examples:

- Category → category (outcome):

  E.g. 'only 15% of nurses are male'

  E.g. '93% of computer programmers in the company are atheists'

- Category → continuous variable (outcome):

  E.g. 'the average pay of tall people is 5% lower than the average pay of short people'

  E.g. 'the average lifespan of Northerners is 5 years lower than of Southerners'

- Category (outcome) → continuous variable:

  E.g. 'the average height of senior executives is 1.82 m but the average height of other executives is 1.84 m'

The argument is that the difference exists, the group membership seems to have no necessary reason for affecting the outcome, and therefore the problem must be unfair bias.

This is not a strong argument.

As with cooperation decisions, seemingly irrelevant characteristics can be statistically relevant. For example, in the hypothetical example about executives of different heights, maybe younger executives tend to be a little taller and that's why the senior executives are a bit shorter on average. Being shorter is not a cause of getting a senior job. Being older is a cause of being shorter and of working longer and getting a more senior job.

Multiple regression

It is better to use a statistical method that considers more than one characteristic simultaneously. This could be a multiple regression that tries to predict the outcomes for individuals. From this it may also be possible to predict the average outcomes for groups.

This does not in itself identify unfair assessments of people but it can help to quantify the extent of overall unfairness, if any, and the impact of unfair assessments will usually be less than this. If there is no unfairness then there probably are no unfair assessments (though something may have been done to compensate for unfair assessments).

The overall idea is to find more and more directly relevant variables that predict outcomes statistically, reducing the predictive power of sensitive variables that may initially have been statistically relevant.

> E.g. Suppose we are trying to understand who gets a senior executive job and who does not. The regression might try to explain this using their height, but also age,

highest level of education, IQ score, years of relevant experience, and anything else that seems likely to help. The importance of height in the regression model is likely to dwindle to nothing when other, more directly relevant variables are included.

If adding more directly relevant variables does not remove the statistical relevance of the sensitive variables then the chances of unfair bias rise.

However, this approach has its problems. It can be hard to get enough data to explain outcomes. The procedure is usually unable to reliably separate out the effects of highly correlated variables, leading to a spill-over that makes irrelevant factors look like they have some relevance. (In general, regressions are much better at *making* predictions than *explaining* their predictions.)

Just because the weight of a variable that is thought to represent unfair bias is not zero does not prove that bias exists. It may just be that a relevant variable has yet to be found and that, when it is, the unfair bias weight will become insignificant.

Decomposition

The factors that lead to overall differences in outcomes are many, and often nothing to do with unfairly biased assessments of people. To identify such bias requires careful analysis to study the effect, if any, of each of many factors.

E.g. Suppose that 80% of speech therapists employed by a hospital are female. In the general population about 51% of people are female. Does this mean that the HR department and others at the hospital have assessed male and female speech therapists unfairly and so recruited too few men?

The comparison with the general population is irrelevant. What would be much more relevant is the percentage of suitably qualified applicants for speech therapy jobs at the hospital who are male and female. (Roughly 97.5% of UK speech therapists are female according to research by Litosseliti and Leadbeater, 2013). Paying attention to that mix helps to screen out the influence of factors outside the hospital, such as the career choices of young people long before they have any contact with the hospital.

In some cases there may be an influence that might have led to unfairly biased assessment but not necessarily.

E.g. Suppose that a young man is told by his school careers advisor that studying psychology is a 'girly' choice and advised to pick computing instead. The young man subsequently chooses to study computing. The advice might have had an influence, but perhaps the young man would have chosen computing anyway, having already noticed that the job prospects are better or because of being a keen amateur coder.

Having detected some unfair bias there is still a need to establish if it is an unfairly biased assessment, or if some other form of bias is involved.

Conclusions

To establish that some assessments are unfairly biased it is necessary to establish that they are biased and that the bias is unfair. Without bias there is no unfair bias.

Consequently, a key part of assessing unfair bias is to understand the truth. If assessments are accurate and if the total set of assessments used is not misleading through being selective then no bias is present. This is true even if the

assessments are negative and particularly negative for a group defined by a legally protected characteristic.

When making assessments of people we should prefer the truth and implement interventions aiming for equity separately from assessment.

A major goal is efficient use of human resources, which involves helping everyone to be as productive and useful as they can be. This is not the same as making everyone equally productive or useful.

References

American National Electoral Studies (2016). 2016 Pilot Study.

https://electionstudies.org/data-center/anes-2016-pilot-study/

Ashton, M. C., & Esses, V. M. (1999). Stereotype accuracy: Estimating the academic performance of ethnic groups. *Personality and Social Psychology Bulletin*, *25*(2), 225-236.

Bergsieker, H. B., Shelton, J. N., & Richeson, J. A. (2010). To be liked versus respected: Divergent goals in interracial interactions. *Journal of personality and social psychology*, *99*(2), 248.

Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied Social Psychology*, 23(23), 1933-1967.

Careersmart website: *Which jobs do men and women do? Occupation breakdown by gender*. https://careersmart.org.uk/occupations/equality/which-jobs-do-men-and-women-do-occupational-breakdown-gender

Cook, C., Diamond, R., Hall, J., List, J. A., & Oyer, P. (2018). *The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers* (No. w24732). National Bureau of Economic Research.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Glick, P., & Fiske, S. T. (1999). The Ambivalence toward Men Inventory: Differentiating hostile and benevolent beliefs about men. *Psychology of women quarterly*, *23*(3), 519-536.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6), 1464.

Johnson, J. D., & Lecci, L. (2003). Assessing anti-white attitudes and predicting perceived racism: The Johnson-Lecci scale. *Personality and Social Psychology Bulletin*, *29*(3), 299-312.

Johnson, D. J., Tress, T., Burkel, N., Taylor, C., & Cesario, J. (2019). Officer characteristics and racial disparities in fatal officer-involved shootings. *Proceedings of the National Academy of Sciences*, *116*(32), 15877-15882.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of personality and social psychology*, 55(6), 893.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *Aea papers and proceedings* (Vol. 108, pp. 22-27).

Leitch, M. (2019). *Making statements about demographic groups*. Available at: http://www.workinginuncertainty.co.uk/groups.pdf

Litosseliti, L., & Leadbeater, C. (2013). Speech and language therapy/pathology:

perspectives on a gendered profession. *International Journal of Language & Communication Disorders*, *48*(1), 90-101.

Maltese, S., Baumert, A., Schmitt, M. J., & MacLeod, C. (2016). How victim sensitivity leads to uncooperative behavior via expectancies of injustice. *Frontiers in psychology*, *6*, 2059.

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology*, *36*(9), 929.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (p. 91–125). Academic Press.

Metcalf, H., & Rolfe, H. (2010). Caste discrimination and harassment in Great Britain. *London: National Institute of Economic and Social Research*.

Nawaz, M. (2012). *Radical: My journey from Islamist extremism to a democratic awakening*. Random House.

ONS (2019). *Internet users, UK: 2019*. Available at: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2019

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European journal of social psychology*, 25(1), 57-75.

Pew Research Centre (2014) *Religious Landscape Study*. Available at: https://www.pewforum.org/religious-landscape-study/religious-tradition/catholic/belief-in-god/

Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men?. *Journal of personality and social psychology*, *87*(4), 494.

Wyness, G. (2016). Predicted grades: accuracy and impact. *London: UCU*. Available at: https://www.ucu.org.uk/media/8409/Predicted-grades-accuracy-and-impact-Dec-16/pdf/Predicted_grades_report_Dec2016.pdf