Confronting bias in judging:
A framework for addressing psychological biases in decision making

Tom Stafford, Jules Holroyd, Robin Scaife
University of Sheffield

June 2018, version 4

ABSTRACT

Cognitive biases are systematic tendencies of thought which undermine accurate or fair reasoning. An allied concept is that of 'implicit bias', which are biases directed at members of particular social identities which may manifest without individual's endorsement or awareness. This article reviews the literatures on cognitive bias, broadly conceived, and makes proposals for how judges might usefully think about avoiding bias in their decision making. Contra some portrayals of cognitive bias as 'unconscious' or unknowable, we contend that things can be known about our psychological biases, and steps taken to address them. We argue for the benefits of a unified treatment of cognitive and implicit biases and propose a "3 by 3" framework which can be used by individuals and institutions to review their practice with respect to addressing bias. We emphasise that addressing bias requires an ongoing commitment to monitoring, evaluation and review rather than one-off interventions.

BIAS AND JUDGES

> "The big problem, as it is everywhere, is with unconscious bias. I dare say that we all suffer from a degree of unconscious bias, and it can occur in all sorts of manifestations. It is almost by definition an unknown unknown, and therefore extraordinarily difficult to get rid of, or even to allow for. "
> - Lord Neuberger, President of the Supreme Court of the United Kingdom, "Fairness in the courts: the best we can do", Address to the Criminal Justice Alliance10th April 2015

Lord Neuberger's comments reflect a growing awareness of the relevance of findings from the behavioural sciences to the practice of judging. In this paper, we aim to show that bias can be addressed, and - if not eliminated from our minds entirely - that there are reasonable steps which professional decision makers, including judges, can take to minimise the negative impacts of bias. Rather than bias remaining an 'unknown unknown', we present psychological research that points towards evidence-informed steps in recognising and dealing with bias.

There is ample reason to worry that judges and court judgements can be biased. Studies of the US court system have found that multiple stages in the prosecution of offences are influenced by non-legal factors (Rachlinski & Wistrich, 2017). One important example is the influence of offender ethnicity on judge's decision (Kutateladze et al, 2014). This includes sentencing decisions by judges (Burch, 2015; Spohn and DeLone, 2000; Hester and Hartman, 2017). The racial disparity in sentencing remains even when legal factors, such as crime severity or the extent of the defendant's criminal history are taken into account (Rehavi and Starr, 2014).

Experimental studies have shown that judge's automatic associations are in line with those of the general population – i.e. often contaminated by prejudicial associations (Rachlinski et al, 2009). In contrast to this, 97% of judges surveyed in that same study declared themselves above average in their ability to avoid racial bias. Other experiments have shown that judges (Englich and Mussweiler, 2001), prosecutors (Englich and Mussweiler, 2006), and jurors (Chapman and Bornstein, 1996) are all influenced by non-legal factors in decisions about the size of compensation awards (for example, being swayed by the prosecution's suggested amount, amounts suggested by journalists or even suggested by the random roll of the die – a phenomenon known as the Anchoring Effect; Tversky and Kahneman, 1974).

Other writers before us have dwelt on the reality of bias in the courtroom (Jolls and Sunstein, 2006; Kang et al, 2012; Rachlinski and Wistrich, 2017, Lee 2013, Holroyd & Picinali forthcoming). Our hope in this paper is to, firstly, offer an analysis of *types* of bias, and, secondly, suggest a framework within which to develop strategies to confront these biases.

We briefly review research on psychological biases, their nature and origin, since we believe that appreciating how psychologists view bias is helpful in interpreting this literature in an applied context. We then summarise how researchers into the topic of decision making have concluded that bias can be addressed, and the obstacles to doing so effectively. Finally, we present the rationale for our own bias interventions and a framework judges can use to systematically consider how to address potential bias in their judgements.

1. BIAS IN PSYCHOLOGICAL SCIENCE

In conventional usage ‘bias’ is a bad thing, suggesting unwarranted prejudice or unfair treatment. In general, the legal sense of bias also cleaves to this meaning. To be biased is a failure of impartiality and so, by definition, to produce bad judgements[1].

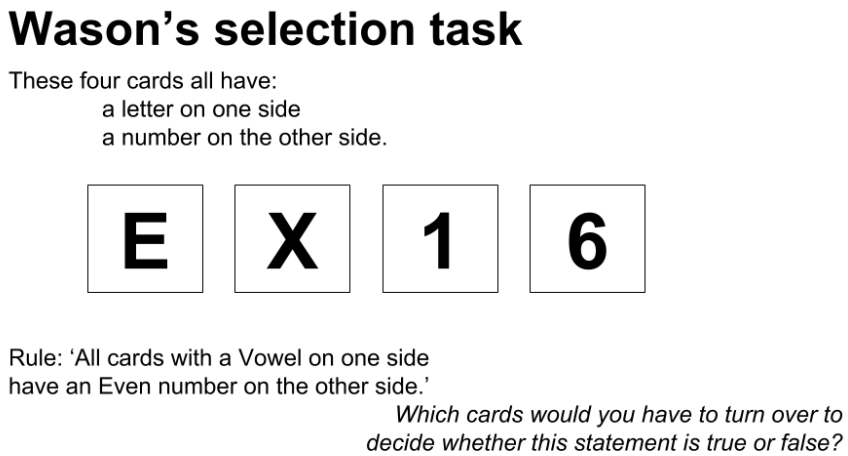
Within psychological science two different research traditions have developed which are relevant to our consideration of bias. Each defines bias somewhat differently, both from each other and from the conventional sense.

Cognitive bias:

One of these traditions is the ‘judgement and decision making’ literature, the concern of cognitive psychologists which focusses on the mechanics of human reasoning, particularly with respect to logic and probability (Kahneman, Slovic and Tversky, 1982). Within this tradition biases are revealed in contrast to the principles of reasoning which they violate; the fallible human reasoner contrasted with the ‘rational actor’ of economic and logical theory which operates in perfect accordance to the principles of those domains.

To illustrate this tradition, let us consider three paradigmatic studies, and the biases which they reveal.

The Wason Selection Task (Wason, 1966, 1968) is a simple reasoning problem with important consequences. The scenario of the task, illustrated in Figure 1, is this: There are four cards, each of which has a single vowel on one side, and a single digit on the other side. As presented, you can see the following on the faces of the four cards: E X 1 6. The following rule is proposed “If a card has a vowel on one side it must have an even number on the other”. The task is to correctly identify which cards you must turn over to check if the rule is true.



1 There is an additional sense in which you can be ‘biased’ by having a conflict of interest. This kind of bias is not about the actual processes involved in judging but in the conditions which must be seen to hold before and after a judgement - namely that an unbiased decision must also seem to be impartial to an objective third party.

Figure 1: The Wason Selection Task

Most participants get this wrong. Typically 80% of participants will select the E and 6 cards to turn over. The correct answer is that the E and 1 cards must be turned over. Turning over the E card tests the rule (if an odd number is discovered, the rule is falsified), but turning over the 6 card can only confirm the rule, if a vowel is found; if a consonant is found then the rule may still be true (the rule is silent on what must be on the other side of cards showing a consonant). This is why the majority conclusion that E and 6 should be turned over is wrong. The 1 card should be turned over because it affords the possibility of falsifying the rule (if a vowel is found). Accordingly, it is only necessary to inspect the E and 1 card, since these are required to confirm or falsify the rule. The common instinct to inspect the card which can only provide evidence in support of the proposition (here, the '6' card) can be taken as an example of confirmation bias, an unfortunately ubiquitous feature of human reasoning to seek out evidence that supports what we already believe (Nickerson, 1998). The assumption is that a common intuition is to accept the rule, reasoning as if it were true and seeking out evidence that might confirm it.

Depending on the precise relationship between the specificity of our beliefs and the evidence available which bears on them, this is a failure of rational information seeking, neglecting the scientific principle that the strongest theories are those which resist falsification rather than accrue confirmation (Popper, 1959). Note however that it is not irrational to seek confirming evidence, only irrational to make insufficient efforts in seeking disconfirming evidence.

The Linda Problem (Tversky and Kahneman, 1982,1932), illustrates a bias in our instincts concerning probabilities. Here is the original problem, with the 1980s wording, when both bank tellers and fear of nuclear war were more common:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

Figure 2: The Linda Problem (conjunction fallacy)

Over 80% of participants answer that (2) is more probable. This violates the rule of probability which states that the probability of A and B must be less than or equal to the probability of A alone (so p(Linda is a bank teller) ≤ p(Linda is a bank teller).p(Linda is active in the feminist movement). Since option 2 is more specific than option 1 it must, strictly, be less – not more - likely, This error is known as the conjunction fallacy, as was explained by the original researchers as reflecting our tendency to base probability judgements on how representative or prototypical an outcome appears.

Our third example of a cognitive bias is known as the Decoy Effect (Huber, Payne and Puto, 1982). We will illustrate the bias using the example described by Ariely (2008). Imagine you can purchase an online-only subscription to The Economist magazine for £59, or a print-only subscription for £125. Many of us, as with the majority of a sample Ariel asked, would prefer the cheap, online only option. The Decoy Effect is that adding a third option, which is less attractive version of one of the options, can alter these preferences. In this case, and based on a real-advert for that magazine, Ariely also tested a version where the options were online only (£59), online and print (£125), and print only (£125). In this scenario the majority preferred the online and print option. Obviously nobody preferred the print only option, the decoy, but its mere presence acted to enhance the attractiveness of the more expensive option due to the contrast. The decoy effect shows that our preferences can be shifted around by seemingly irrelevant options. The bias is only revealed because we know people's preference without the decoy. Note that it isn't possible to say which is the correct choice, only that the majority preference for the cheaper online only option in one case, and for the more expensive online and print option in another seems to be an example of inconsistency.

Ariely uses the Decoy Effect to illustrate the essential relativity of our preferences – that we are forced to evaluate things in their context of other things. It is not hard to imagine judicial scenarios when such biases might operate.

Note some common features of these examples. The errors are concerned with the structure of reasoning demanded by the problems - they are not intended to illustrate something about the content of the problems (not about vowels and even numbers, Linda's feminism or magazine subscriptions). The errors are identified as such because a standard of rationality is violated: either a failure of logical inference, of correct probabilistic inference or of simple consistency.

Research in psychology has used tasks like these to illustrate systematic errors in human judgement and decision making. These systematic errors are, in turn, interpreted as evidence of their generating mechanisms: tendencies of thought, known as heuristics, which may sometimes serve useful ends, but which also produce characteristic errors (Kahnemann, Slovic & Tversky, 1982; Gigerenzer & Todd, 1999). The consensus in this area is that biases are not unnecessary errors which somehow contaminate what would otherwise be a purely rational mind. Rather, the biases are an unavoidable side-effect of the organising principles of the mind (Gigerenzer & Selten, 2002; Simon, 1982), and so the heuristics which they arise from are – in some part - constitutive of thought, rather than extraneous influences. However, since they can sometimes lead us to error, we need to be alert to the possibility of these biases and the occasions on which we may err as a result.

Social bias:

A second tradition in the study of psychological bias is the study of what has been called 'mental contamination' (Wilson & Brekke, 1994), whereby our judgement is influenced by unconscious or unmonitored mental processes. Much of this work focusses on the contamination of judgements of or by other people, and information about the social categories to which they belong.

Examples include experiments which appear to show that people who complete a word-puzzle in which the answers are words associated with being elderly('old','grey','zimmerframe', etc) walk more slowly as they exit the building after they think the experiment is over (Bargh 1996; but also see Doyen et al, 2012; Stafford, 2014); experiments showing that even those who profess anti-racist beliefs find it can find it harder to associate positive words with black faces compared to with white faces (the 'implicit association test'; Greenwald, McGhee & Schwartz, 1998; Nosek, Greenwald & Banaji, 2007); and experiments which show that candidates with identical CVs are rated as less competent and worthy of lower starting pay if a woman's name is put on them rather than a man's (Moss-Racusin, 2012; see Jost et al 2009 for an extensive summary of research into these kinds of biases).

Certain instances of mental contamination that involve stereotypes or evaluations of social groups have been referred to as 'implicit bias' (Holroyd, Scaife & Stafford, 2017; Greenwald & Krieger, 2006; Jolls & Sunstein, 2006). The contrast is with explicit bias, a prejudice which one is aware of and endorses. An implicit bias is revealed in prejudicial actions which may contradict your endorsed views. This may be in subtle aspects of behaviour, such as differential warmth in body language when interacting with people from minority ethnic backgrounds (Word, Zanna & Cooper, 1974; Dovidio, Kawakami & Gaertner, 2002), or more serious divergences in treatment such as differential likelihood of falsely identifying a young man as carrying a weapon if he is black rather than white (Correll et al, 2002). The experiments reported by Word, Zanna and Cooper (1974) are particularly instructive with regard to implicit bias. In this study white participants were recruited under the ruse of an investigation into group decision making processes. Participants were asked to interview applicants for a position on their team. If the applicant was black the study participants sat further away from them, leaned forward less in conversation and showed fewer signs of positive engagement (such as nodding of smiling). Follow up work has confirmed this result and showed that participants may not be aware that their behaviour is being influenced by the applicant's race in this way (Dovidio, Kawakami and Geartner, 2002). Crucially, the original study also investigated the effect that negative social signals can have on interview performance. For this second experiment, the authors specifically trained confederates to treat participants with different levels of positivity – so that for some they mirrored the reduction in positive social signals that participants displayed in the first experiment. When treated in this way, job applicants sat further from the interviewer, made more speech errors and spent less time answering the interview questions. When their interview performance was evaluated by independent judges they were judged less adequate for the job. The study authors offered their investigation as an example of a self-fulfilling prophecy. Black candidates are treated with less warmth, and so find it harder to perform well in interviews and are judged as less competent, which perpetuates some part of the stereotype which drives interviewer attitudes.

Combining research into bias for practical action

Note that cognitive bias is defined with respect to standards of logic and mathematics which are relatively easy to articulate (even if their application to everyday human action is not straightforward). In contrast, defining bias with respect to explicit endorsement, recognition or intention means it is less clear which moral or epistemic principles are being violated. We do not wish to take a position on the rationality or adapative value of possessing implicit biases (Picinali 2016; Gendler, 2011). What we do wish to claim is that the study of cognitive biases can inform our treatment of implicit or (so-called) 'unconscious' bias.

The key points that emerge from these bodies of research are as follows: first, that our susceptibility to certain kinds of biased reasoning is pervasive, and may not always be undesirable: whilst cognitive biases may sometimes cause us to make irrational judgements, they may otherwise be useful cognitive devices on which we often rely. Second, however, this reliance on quick heuristic devices can be extremely problematic in social contexts in which stereotypes and misconceptions attached to different social identities prevail. Moreover, whilst the tendency to rely on heuristics may be unavoidable, their particular content (namely, the specific stereotype or association), that fills out the fast cognitive links we make, seems to be highly malleable and influenced by social context. This presents us with a problem: given our cognitive dispositions, how should we shape our environments and institutions such that problematic biases do not populate our minds or influence our decisions and actions, and so that the biases are mitigated, or else align with our endorsed values?

Another benefit of considering both cognitive and social biases is that for an individual decision maker, it is not possible to divide one's thinking about biases neatly between two types. Not only must a judge guard against both cognitive and social biases - against both rational and moral risk in judgements - but there is the worry that the two types of bias may combine. So, for example, an initial impression of an appellant may be contaminated by their social identity with respect to ethnicity, sex, etc and that initial impression may feed a tendency to confirmation bias in the way matters of fact are pursued.

Looked at from the other direction, some evidence of bias in court judgements shows divergent outcomes according to factors such as ethnicity or gender, but it tends not to reveal *how* such factors distorted judgements (Burch, 2015; Spohn and DeLone, 2000; Hester and Hartman, 2017). Bias in sentencing, for example, may arise because a judge faces biased witnesses or officials, rather than the judge his or her-self expressing bias directly (cf Silbershan et al, 2017). The moral from research into biased decision making is that there are many more routes to a prejudicial judgement than direct translation from a prejudicial attitude held by a single individual.

With these caveats in mind, we turn now to the essential issue – once we are convinced of the risks of making biased judgements, what can we do about it?

2. WHY BIAS MITIGATION IS HARD

There are a number of challenges for a judge seeking to make fair decisions without discriminating against individuals due to their membership of a particular social group, or due to idiosyncrasies of reasoning, or some combination of both. Before we review specific strategies and their effectiveness we will introduce important background factors which, we believe, set reasonable expectations on how difficult a task 'de-biasing' is.

Wilson & Brekke (1996) present a helpful process diagram which makes explicit the necessary conditions for removing unwanted bias from a decision (Figure 3). In short: not only must you be aware of the bias and motivated to correct it, but you must know the direction and magnitude of the distortion(s) it has produced on your judgement, as well as possessing the means to correct those distortions. Being aware of your own biases is hard enough (Pronin, 2007), it is an extra step to be fully aware of the exact distortion they introduce into our judgements.

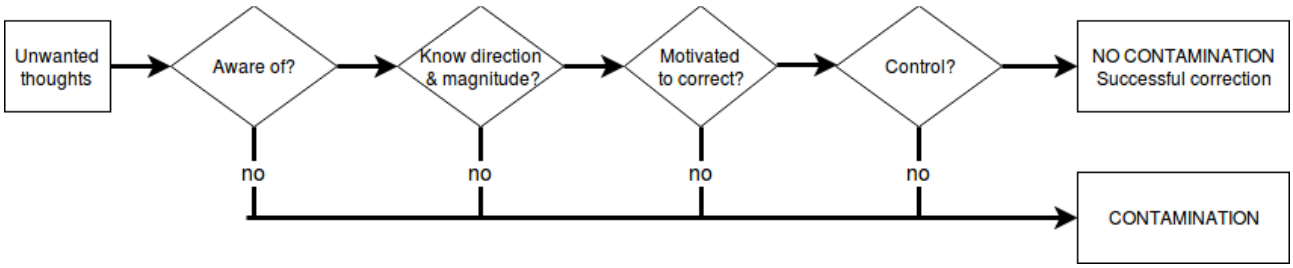


Figure 3. Redrawn from Wilson & Brekke (1994), Figure 1 See https://figshare.com/articles/_The_process_of_mental_contamination_and_mental_correction_after_Wilson_Brekke_1994_Figure_1/4233215

Many biases are deeply entrenched, either due to social experience, direct or vicarious, or due to the way the architecture of our minds interacts with our environment (Anderson 2010, p.51). So, as an example of the first kind, the automatic association of men with leadership is one that anyone growing up in our culture is likely to make, given a lifetime's exposure to a preponderance of leadership positions being filled with men and the interaction of culture and language that tends to reflect that preponderance (such that the traditional name for the leader of a committee is chairman and myriad other examples). As an example of the second kind, there is a well-known phenomenon whereby other things being equal, we tend to prefer the familiar. This is a bias if there is no logical reason to prefer the familiar option, but it makes sense if we consider the likely costs and benefits of a "go with what you know" strategy across our lifetimes, or across evolutionary time. The problem for judges is that a preference for the familiar, whilst adaptive for our ancestors during the dominant conditions of the evolution of the human mind, is not a legitimate factor for a judge or juror when weighing which witness to believe evidence from, for example.

The forces which produce bias may be deeply entrenched, but that doesn't mean that they can't be tempered, or the decisions themselves protected from bias, but it does suggest that simple fixes are likely to be of limited effectiveness.

Prior research has shown that some obvious strategies are notable for being minimally effective or even prompting backfire effects, at least when employed on their own without being embedded in a broader anti-bias framework.

Awareness raising: Raising awareness of prevalence of stereotyping can inadvertently carry the message that "everyone is doing it". There is some evidence that this can undercut and even increase the likelihood that individuals will allow stereotypes to influence their judgement (Duguid & Thomas-Hunt, 2015). Important relevant work has shown that social norms have a powerful influence on behaviour, and warning messages can inadvertently advertise an undesirable social norm (Cialdini, 2003). So, for example, an intervention that warns "most CEOs unfairly believe that women are ineffective leaders" is advertising both the unfairness of the belief, but also its pervasiveness. This may have inadvertent impact in legitimising behaviours or beliefs.

Suppression: Simple injunctions not to be prejudiced, or to ignore social categories such as race can also be ineffective or backfire (Paluck & Green, 2009; Legault et al, 2011; Apfelbaum et al, 2008). It isn't clear that the most pernicious biases are under an individual's direct control, so we might expect the injunction to avoid being prejudiced to be ineffective. Further, the feeling of objectivity may itself allow inadvertent expressions of prejudices (Uhlmann & Cohen, 2005,2007). Moreover, strategies of making race salient, rather than aiming for 'colour-blinded' approaches, better enable individuals to identify and combat racial biases (Lee 2013).

Targeting implicit associations directly: Although there have been efforts around interventions to shift specific measures of implicit bias (e.g. Lai et al, 2014, 2016; Devine et al, 2012), it is not clear how these might translate into practical applications (for example, some of the successful interventions are based around faking scores on the a specific and well known measure of implicit attitudes, the IAT). Further, the IAT itself has come under fire as being of dubious relevance to ecologically valid instances of discrimination (Oswald et al, 2013; Greenwald et al, 2009), especially when compared to explicit measures. Finally, whilst implicit biases may be malleable, meta-analyes indicate that changes in implicit biases do not necessarily translate into changes in behaviour (Forscher et al 2017).

Targeting prejudicial attitudes directly: A general review of prejudice reduction strategies concluded that widespread weaknesses in research methodology leave the effectiveness of many interventions unknown (Paluck & Green, 2009). This included workplace diversity training. Training of managers, as opposed to general employees, on diversity issues has also been found to be among the least effective of interventions (Kalev et al, 2006). This said, Bezrukova et al (2016) support the efficacy of diversity training in general, without a specific focus on bias or implicit bias. In other words, it can alter people's beliefs, awareness and professed attitudes, but without guarantee that this will affect the display of bias that concerns us here.

3. EVIDENCE ON ANTI-BIAS STRATEGIES

Against the background of these considerations we now summarise the lessons from research into removing or mitigating bias.

Limits of individual-only interventions

There are parallel seams of research into how to address bias. Reviews which have focussed on the cognitive biases are generally more optimistic about the possibility of removing bias from decisions (Soll et al, in press; Milkman et al, 2009; Wilson & Breke, 1994; Larrick, 2004; Lilienfeld et al, 2009). By focussing on specific cognitive biases which relate to identifiable flaws in reasoning, it has proved possible to successfully employ specific correctives (Morewedge et al, 2015). So, for example, if participants are ignorant of the rules of probability they can be told them, or can be made aware of the tendency to seek confirmatory evidence. Within the narrow scope of specific problems these measures may be enough. Indeed judges in criminal cases will be familiar with the need to instruct a jury in relation to specific points. For example the Turnbull warning is a well-known directive that counsels jurors on the weighting of eye witness testimony, given their tendency to inflate the credibility and import of such evidence.

Reviews of the literature on social bias overlap with the literature on prejudice, discrimination and diversity training, and have tended to emphasise the pervasive and intractable nature of bias. It is beyond the scope of this paper to address all interventions aimed at reducing *explicit* prejudice. Nonetheless, this literature is valuable since it addresses a less narrow problem than the quite specific errors which are the focus of the literature on cognitive bias. Recent reviews of bias interventions have clarified the elements an intervention must get right as a precondition to success. Kalev et al (2006) review corporate affirmative action and diversity policies and conclude that management training was least effective. Structural changes in the form of establishing organisational responsibility, and action plans for dealing with diversity issues, or reducing social isolation of minorities via means such as mentoring programmes were better supported. Paluck & Green (2009)'s review on prejudice reduction concluded that there was moderate evidence for the efficacy of contact between social groups reducing prejudice - again affirming that structural changes, such as increased recruitment from minority backgrounds, are more effective at changing attitudes.

The lesson we take from the literature on social biases is that a full treatment of bias in decision making requires attention to interpersonal, procedural and institutional actions, as well as individual psychology. Although the literature on bias in psychology has a very individualistic frame of reference, the evidence is that individuals are in general poorly situated to be aware of, control and or measure the extent of their own biases (see above; Wilson & Brekke, 1996; Scaife, in preparation). Further, some straightforward individual interventions are of low effectiveness, especially for 'social biases' (e.g. raising awareness, attempting to suppress bias, as discussed above).

A focus solely on individuals puts us at risk of neglecting wider collective responsibility for bias (Dixon et al, 2012), as well as falsely implying that the goal of de-biasing is to solely to purify individuals' cognitions, rather than the goal being wider social aims, such as fairness of procedure or of representation.

Attribution, accusation, blame and accountability

One risk with bias interventions is that they will create backlash effects, whereby individuals feel blamed and respond by rejecting the principles or practices of the intervention. Moss-Racusin et al (2014) emphasise the importance of deploying interventions in a context of common purpose (e.g. "we all want to improve decision making"). Although the risk of a bias intervention being rejected isn't to be treated lightly, recently our work has found that people accused of harbouring implicit biases do not automatically respond by rejecting an anti-prejudicial message (Scaife et al, under review, See also, Czopp, Monteith, & Mark, 2006). This is some cause for optimisim, since it suggests that there is no automatic backlash or rejection effect which accompanies demonstrating to people the reality of their implicit biases.

We have found it helpful to frame the discussion of bias in terms which are both collective, and "forward-looking": "how might *we* ensure that future decisions are fair and impartial?", rather than to focus on the attribution of blame for past or hypothetical decisions (see Watson, 1996; Zheng, 2016). The psychological perspective on bias prioritises a backward-looking and attributive approach - we seek to identify which biases, operating in which individuals, generate biased outcomes. A forward looking perspective is more compatible with an aspiration to collective action, based on a shared moral purpose. It also encourages a focus on those constructive actions that can lessen the possibility or impact of bias, rather than on questions of attribution which are both practically and conceptually hard to answer.

Need for ongoing review

Moss-Racusin et al (2014) set out a framework for scientifically informed diversity interventions. They observe, as reported by Paluck & Green (2009), that most diversity interventions rely on lecturing, despite the well-known superiority in effectiveness of active learning strategies that promote engagement with course content.

A report of just such a scientifically informed intervention (Moss-Racusin et al, 2016) emphasises the need for any intervention to leave participants "action-ready", with a focus on what they may positively do to achieve a goal (e.g. treat people fairly) rather than merely with the injunction to avoid a negative outcome (e.g. display bias).

In parallel to the need for any training to leave participants ready to engage in positive action, any anti-bias strategies should be combined with monitoring, review and evaluation. A single training session is inadequate to address bias either within individuals or organisations.

Our consideration of the nature of social biases puts into perspective the limited effectiveness of many interventions. If these biases are due to automatic associations we make between social categories, built up over a lifetime of direct and vicarious experience, then it isn't reasonable to expect an hour long intervention to overturn them. An analogy we find helpful is that of the relationship between dieting and a healthy meal. If we were overweight we wouldn't eat one healthy meal and think "right, I'm fit now". Instead, a healthy weight would require an ongoing commitment to eat healthy meals in the future, combined with broader lifestyle changes. Similarly with anti-bias strategy, learning about bias and thinking about it for an hour, is not a recipe for the effective ongoing change. Therefore, our belief is that any successful intervention must have built in mechanisms to encourage follow up engagement.

4. A “3 by 3” FRAMEWORK FOR ADDRESSING BIAS

We conclude by introducing a 3x3 framework which we have used with judges to review their approach to decision making (Figure 4). The framework provides a structure within with which to record and explore different anti-bias strategies in decision making. Rather than be a substitute for any individual anti-bias strategy, it is intended to help individuals and institutions recognise and organise existing aspects of their practice which address bias, as well as to reveal where gaps exist that may be filled.

The first dimension of this framework is to categorise strategies according to *how* they work - whether they mitigate against bias (reduce the impact of bias on outcomes, whilst potentially leaving any bias unaffected) or insulate against bias (by removing bias-triggering information, or otherwise preventing any bias from operating in a decision) or remove a bias from individual’s cognitions (de-biasing proper).

The second dimension of this framework is to categorise strategies according to *where* they work - whether the strategies focus on individual action, interpersonal interactions, or institutional changes. This makes explicit the conclusion from existing literature on addressing bias that strategies which only focus on individuals are, if deployed in isolation, an inadequate subset of all available strategies.

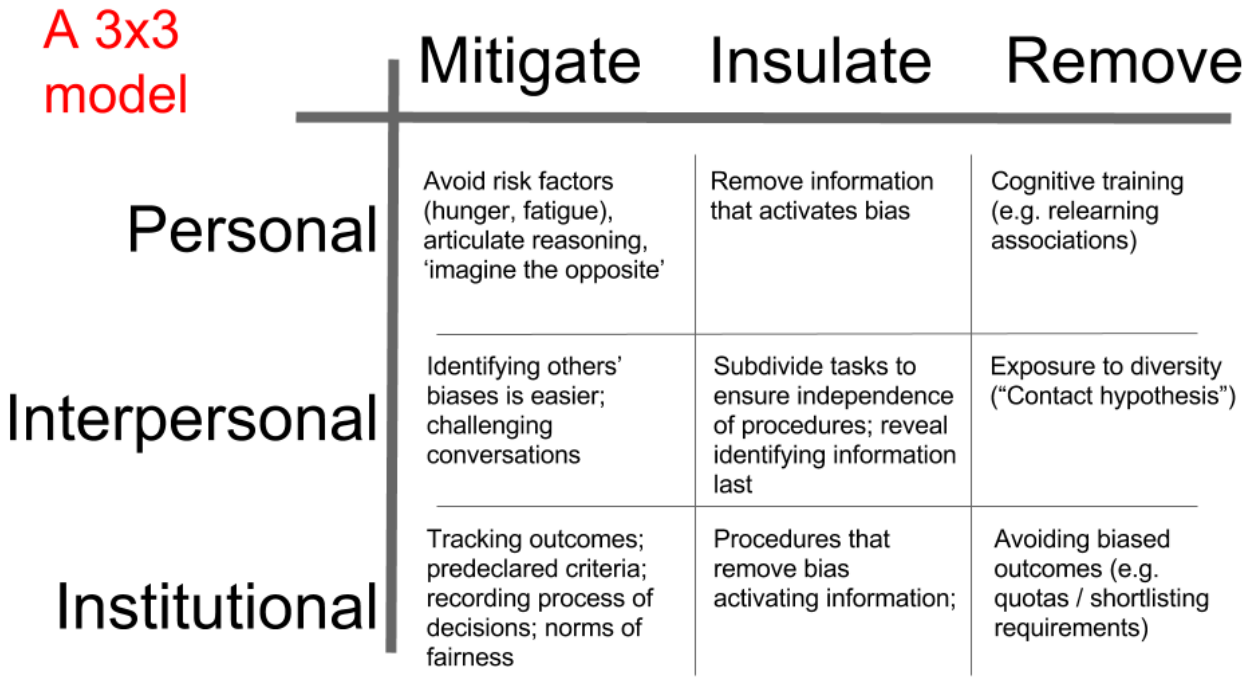


Figure 4: A 3x3 framework for addressing bias, with possible strategies as illustration.

Combining these two dimensions we get a 3x3 framework, shown in Figure 4, populated by illustrative strategies. These illustrative strategies are not meant to be definitively positioned, or exhaustive. The first key task for anyone thinking about bias is to identify strategies of their own that they already deploy and to think where those strategies might be appropriately located within this framework. Appendix 1 provides a review of strategies which have been previously discussed in the literature. The second key task is to identify how the effectiveness of that strategy could be evaluated.

The levels of each dimension have different qualities to recommend them. Insulating a decision from bias may have greater effectiveness than trying to mitigate bias, but both leave the existence of bias unaffected in the longer term. Strategies aimed at removing bias may be more desirable in the longer term, if it is conceivable that such a bias can be removed (which it may be for something like a bias against women leaders, for example, but perhaps not for something which is part of our cognitive machinery like a tendency to confirmation bias). We have already commented on the difficulties for individuals to recognise and accommodate their own biases, which is why we wish to emphasise the range of strategies that also exist at the interpersonal and institutional level. Obviously many strategies will breach levels. For example, a strategy such as a policy or procedure enacted at the institutional level will have to be considered and adopted by specific individuals.

As well as explaining how we think about bias, our intervention also asks participants to consider some paradigmatic cases of bias in order to encourage reflection on their own decision making and sharing of strategies among the attendees. This reflects the finding that active learning is a more effective technique than lecturing, in the context of bias training programmes.. We provide these case-studies, each of which is based on a pivotal psychology experiment on bias, in the online supplementary material (https://osf.io/cfgh6/). Also provided is an annotated reading list of resources about bias, with a focus on general introductions and over-views, as well as more specific pieces relevant to legal professionals

5. CONCLUSION

Previous literature has focussed on establishing the reality of bias, both within the legal professions and in the general population. Psychologists have been preoccupied with exploring the potential psychological mechanisms for generating biased outcome. Now, we believe, is the time to draw together the wealth of research in this area with a focus on practical steps for addressing bias. Far from being an “unknown unknown” research has revealled much about the nature of bias and suggested practical steps for addressing it. Unbiased decision making is a lofty but necessary goal. Although there is no single action or innovation which guarentees protection against our psychological biases, there are collective and individual actions which can render us less vulnerable.

6. REFERENCES

Apfelbaum, E., Sommers, S., & Norton, M. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. Journal of Personality and Social Psychology, 95, 918-932

Ariely, D. (2008). Predictably irrational. New York: HarperCollins.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation.

Burch, T. (2015). Skin Color and the Criminal Justice System: Beyond Black-White Disparities in Sentencing. *Journal of Empirical Legal Studies*, *12*(3), 395–420. https://doi.org/10.1111/jels.12077

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current directions in psychological science*, *12*(4), 105-109.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. Journal of Personality and Social Psychology, 83, 1314-1329.

Cuthbert, L. (2015). Too Confident By Half. Tribunals, 10.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: reducing bias through interpersonal confrontation. Journal of personality and social psychology, 90(5), 784-803

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences, 108(17), 6889-6892.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. Journal of experimental social psychology, 48(6), 1267-1278.

Dixon, J., Levine, M., Reicher, S., & Durrheim, K. (2012). Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution?. *Behavioral and Brain Sciences*, *35*(06), 411-425.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. Journal of personality and social psychology, 82(1), 62-68

Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Context management toolbox: a linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. Journal of forensic sciences, 60(4), 1111-1112.

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. Journal of Applied Psychology, 100(2), 343-359. http://dx.doi.org/10.1037/a0037908

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A.
Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 422-444). Cambridge, UK: Cambridge University
Press.

Gawande, A. (2009). The checklist manifesto: How to get things right. New York: Metropolitan Books.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. Philosophical Studies, 156(1), 33-63.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gigerenzer, G., & Selten, R. (2002). Bounded rationality: The adaptive toolbox. MIT press.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. BMJ: British Medical Journal, 741-744.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. California Law Review, 94(4), 945-967.

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), 1464-1480

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of personality and social psychology, 97(1), 17.

Green, D.M., Swets J.A. (1966) *Signal Detection Theory and Psychophysics*. New York: Wiley.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: motivated reasoning and rationality. PSYCHOLOGY OF LEARNING AND MOTIVATION, VOL 61, 61, 41-102.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. Psychological Science, 20(2), 231-237.

Hester R. & Hartman T.K. (2017) Conditional Race Disparities in Criminal Sentencing: A Test of the Liberation Hypothesis From a Non-Guidelines State. *Journal of Quantitative Criminology, 33*(1), 77-100

Holroyd, J., Scaife, R., Stafford, T. (in press). What is Implicit Bias? Philosophy Compass.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. Journal of Consumer Research, 9(1), 90–98. https://doi.org/10.1086/208899

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. Journal of Personality and Social Psychology, 69, 1069-1086.

Jolls, C., & Sunstein, C. R. (2006). Debiasing through law. *The Journal of Legal Studies*, *35*(1), 199-242.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management science*, *39*(1), 17-31.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American sociological review*, *71*(4), 589-617.

Kang, J., Bennett, M. W., Carbado, D. W., Casey, P., Dasgupta, N., Faigman, D. L., ... & Mnookin, J. (2012). Implicit bias in the courtroom. UCLA Law Review, 59(5).

Klein, G. (2007). "Performing a Project Premortem". Harvard Business Review. 85 (9): 18–19.

Koehler, J. J., & Meixner, J. B. (2013). Decision Making and the Law: Truth Barriers. Wiley-Blackwell Handbook of Judgment and Decision Making, Forthcoming, 13-04.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of personality and social psychology, 77(6), 1121--1134.

Kutateladze, B. L., Andiloro, N. R., Johnson, B. D., and Spohn, C. C. (2014). Cumulative disadvantage: Examining racial and ethnic disparity in prosecution and sentencing. Criminology, 52(3), 514-551.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., ... & Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. Journal of Experimental Psychology: General, 143(4), 1765.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., Hu, X., McLean, M. C., Axt, J. R., Asgari, S., Schmidt, K., Rubinstein, R., Marini, M., Rubichi, S., Shin, J. L., & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. Journal of Experimental Psychology: General, 145, 1001-1016

Larrick, R.P. (2004). Debiasing. In D.J. Koehler & N. Harvey (Eds.), Blackwell handbook of judgment and decision making (pp. 316–337). Oxford, United Kingdom: Blackwell.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages how motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, 0956797611427918.

Lerner, J.S., & Tetlock, P.E. (1999). Accounting for the effects of accountability. Psychological Bulletin, 125, 255–275.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? Perspectives on Psychological Science, 4, 390-398.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. Journal of personality and social psychology, 47(6), 1231.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and brain sciences, 34(02), 57-74.

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? Perspectives on Psychological Science, 4, 379-383.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. Psychological review, 115(2), 502-517

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2372732215600886.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences, 109(41), 16474-16479.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2014). Scientific diversity interventions. *Science*, *343*(6171), 615-616.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2016). A "Scientific Diversity" Intervention to Reduce Gender Bias in a Sample of Life Scientists. *CBE-Life Sciences Education*, *15*(3), ar29.

Fairness in the courts: the best we can do: Address to the Criminal Justice Alliance. Lord Neuberger, 10 April 2015
https://www.supremecourt.uk/docs/speech-150410.pdf

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. Personality and Social Psychology Bulletin, 26, 1142–1150.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175-220.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. Automatic processes in social thinking and behavior, 265-292.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. Journal of Personality and Social Psychology, 105(2), 171–192. http://doi.org/10.1037/a0032734

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology*, *60*, 339-367.

Picinali, F. (2016). Base-rates of Negative Traits: Instructions for Use in Criminal Trials. *Journal of Applied Philosophy*.

Popper, Karl (1959). The Logic of Scientific Discovery. New York, NY: Basic Books.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, *11*(1), 37-43.

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2008–2009). Does Unconscious Racial Bias Affect Trial Judges. *Notre Dame Law Review*, *84*, 1195–1246.

Rachlinski, J. J., & Wistrich, A. J. (2017). Judging the Judiciary by the Numbers: Empirical Research on Judges

Rehavi, M. M., and Starr, S. B. (2014). Racial Disparity in Federal Criminal Sentences. Journal of Political Economy, 122(6), 1320-1354.

Simon, H. A. (1982). Models of bounded rationality: Empirically grounded economic reason (Vol. 3). MIT press.

Soll, J., Milkman, K., & Payne, J. (in press). A user's guide to debiasing. In G. Keren & G. Wu (Eds.), Wiley-Blackwell handbook of judgment and decision making. New York, NY: Blackwell.

Spohn, C. and DeLone, M. (2000). When does race matter?: An analysis of the conditions under which race affects sentence severity. S*ociology of Crime Law and Deviance, 2*, 3-37.

Stafford, T. (2014). The perspectival shift: how experiments on unconscious processing don't justify the claims made for them. *Frontiers in Psychology, 5*, 1067. doi:10.3389/fpsyg.2014.01067

Stafford, T. (2015) For argument's sake: evidence that reason can change minds. Kindle Direct Publishing

Tversky, A. and Kahneman, D. (1982) "Judgments of and by representativeness". In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

Uhlmann, E. L., & Cohen, G. L. (2007). “I think it, therefore it’s true”: Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*(2), 207-223.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria redefining merit to justify discrimination. *Psychological Science*, *16*(6), 474-480.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), New horizons in psychology I (pp. 106–137). Harmandsworth: Penguin.

Wason, P. C. (1968). Reasoning about a rule. Quarterly Journal of Experimental Psychology, 20(3):273-281.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227-248.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin*, *116*(1), 1171-142.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of experimental social psychology*, *10*(2), 109-120.

Zheng, R. (2016). Attributability, Accountability, and Implicit Bias. In Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics, Jennifer Saul and Michael Brownstein (eds.) New York: Oxford University Press, 2016. 62-89.

ACKNOWLEDGEMENTS

Phil Rostant, Christa Christensen, Barry Clarke and all those judges who took part in training on bias. Paul Cruthers for feedback on a draft. Tara Lai Quin and postgraduate students in the University of Sheffield School of Law who discussed these issues. Robin Zheng and all those who attended the Philosophy of Implicit Bias workshops in Sheffield 2014-2017. The Leverhulme Trust for their support via a project grant on “Bias and Blame” (RPG - 2013-326).