The University Of Sheffield.

This is a repository copy of *The social psychology of discrimination*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/130466/

Version: Accepted Version

**Book Section:**
Holroyd, J.D. (2017) The social psychology of discrimination. In: Lippert Rasumussen, K., (ed.) Routledge Handbook on the Ethics of Discrimination. .

**Reuse**
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

**Chapter 32**

# The social psychology of discrimination

Jules Holroyd, University of Sheffield

Word count:

Abstract

How, if at all, do the findings of social psychology impact upon philosophical analyses of discrimination? In this chapter, I outline key findings from three research programs from psychology – concerning in-group/out-group favoritism; implicit bias; and stereotype threat. I argue that each set of findings presents challenges to how philosophical analyses of group discrimination are formulated, and propose possible revisions to be explored in future work.

## 1. Introduction

Research in social psychology focuses on the psychological mechanisms involved in and underpinning discrimination. The research program is vast (for accessible overviews, see Ramiah et al 2010, for research on gender discrimination in particular see Valian 1999, on racial discrimination see Blank et al 2004). In this chapter I focus on three areas of research from social psychology that have been taken up by philosophers: in-group/out-group favoritism; implicit bias; and stereotype threat. Whilst each has garnered some philosophical attention, little attempt has been made to tease out the implications of this research for philosophical analyses of discrimination. In this chapter I explore the some of the ways in which insights from research from social psychology may require revision or fine-tuning to philosophical analyses of discrimination. Philosophers have good reason to attend to the findings of social psychology: not only to gain an understanding of the mechanisms that may underpin discrimination; but also to inform philosophical analyses of discrimination and ensure they are suitably formulated to capture the full range of the phenomena.

## 2. Discrimination

In order to evaluate the implications of this research for analyses of discrimination, it will help to have an understanding of discrimination in view. For present purposes, we will focus on the analysis offered by Lippert-Rasmussen (2013), which has been recently elaborated and defended. The conditions are intended to capture the notion of group discrimination, which is arguably at issue in common usage of the notion of discrimination. Lippert-Rasmussen proposes the following set of conditions for group (direct) discrimination:

> An agent, X, discriminates against someone, Y, in relation to another, Z, by Φ-ing (e.g. hiring Z rather than Y) if and only if:
>
> i) There is a property, P, such that Y has P or X believes that Y has P, and Z does not have P or X believes that Z does not have P,
>
> ii) X treats Y worse than he treats or would treat Z by Φ-ing, and
>
> iii) It is because (X believes that) Y has P and (X believes that) Z does not have P that X treats Y worse than Z by Φ-ing (15)
>
> iv") P is the property of being a member of a certain socially salient group (to which Z does not belong) (26)
>
> v) Φ is a relevant type of act, policy or practice, and there are many acts etc. of this type, and this fact makes people with P (or some subgroup of these people) worse off relative to others, or Φ is a relevant type of act etc., and many acts etc. of this type would make people with P worse off relative to others, or X's Φ-ing is motivated by animosity towards or dislike of individuals with P or by the belief that individuals who have P are inferior or ought not to intermingle with others (2013: 28)

Lippert-Rasmussen clarifies that conditions i-iii are part of an analysis of generic discrimination (15-22) - a broad notion of discrimination that captures differential treatment of persons such as

exclusion of unqualified candidates from consideration for a job, or judicial conviction of individuals found to be guilty (15). The addition of conditions iv") and v) narrows down the analysis to one that captures those cases in which individuals are differentially treated on the basis of membership of socially salient group, where social salience is a matter of perceived membership being important to the structure of social interactions in a range of social contexts (e.g. race, gender, age) (30). This idea is accommodated by condition iv"). Finally, condition v) of Lippert-Rasmussen's analysis is needed in order to give an analysis that excludes one-shot, unproblematically motivated differential treatment that does not, or would not, constitute a pattern of disadvantaging treatment (28-30). Accordingly, condition v) captures the idea that it is differential treatment which involves actions, or patterns of action, that do or would disadvantage, or which express hostility or judgments of inferiority, with which we are concerned.[1]

We should note that it is this idea of *group* discrimination with which much research in social psychology has been concerned: the driving motivation behind much of the social psychological research is to understand discrimination and the mechanisms underpinning it as it tracks and structures our social identities. Ramiah et al (2010) motivate their field of research by observing that 'the pervasiveness of discrimination and its systematic, and often subtle, expression shapes society in ways that perpetuate inequities' (90). This much indicates that at stake for these authors is discrimination that tracks social identities taken to be important and with significance for the ways in which society is structured.

Note, moreover, that our focus for the most part in this paper is on *direct discrimination*: cases in which an individual (or organization) treats another person or group disadvantageously, because of their social group membership (race, age, gender and so on). This is because much of the psychological research is relevant to direct discrimination, since it speaks to the cognitive mechanisms underpinning differential treatment in this sense. I set aside, for the most part, the important related phenomenon of *indirect discrimination*, which concerns cases where a policy or

procedure is on the face of it neutral, but in fact disproportionately disadvantages members of a particular social group.

With this analysis on the table, and a sense of its relationship to the concerns of social psychological research, we can examine some of the research programs which focus on the cognitive mechanisms that underpin discriminatory judgments and behavior. We will then be in a position to consider what revisions or refinements, if any, this understanding of group discrimination may benefit from.

**3. Discrimination involving In-group favoritism**

Social psychologists have identified in-group favoritism as one of the mechanisms by which differential treatment is perpetrated. So called 'categorization effects' - whether individuals are categorized as 'in-group' or 'out-group' members - have been shown to impact on a wide array of behaviors, from what is credited to the an agent, to the sorts of benefits allocated to her.

For example, various studies have prompted individuals to consider whether behaviors or outcomes, which may have been positive or negative, were the result of an agent's traits (e.g. her hard work or talents), or due to factors external to the agent (such as luck or other people's actions). These studies have focused on different in-group/out-group identities, with Hindu participants being asked to evaluate the described behaviors of Hindu (in-group) or Muslim (out-group) individuals (Taylor & Jaggi 1974); Israeli and Arab American students being asked to evaluate actions from other Israeli and Arab agents (Rosenberg & Wolfsfeld 1977); and even individuals with preferences for Super Bowl teams (Dallas Cowboys vs Pittsburgh Steelers) being asked to evaluate the causes of events involving their respective teams in a game (Winkler & Taylor 1979).

In these studies, the findings indicated that when individuals evaluated out-group members, negative behaviors or outcomes were more frequently attributed to failings on the part of the out-group agent (bad character, lack of effort, and so on). Meanwhile, positive outcomes were explained by situational factors, thereby removing credit from the out-group individual. The

converse pattern was found when participants made evaluations of in-group members: positive outcomes were more frequently attributed to traits of the agent, whilst negative actions or outcomes were attributed to situational factors out of the agent's control.[2]

In another study, participants viewed video recordings of behaviors (a 'shove') that could be interpreted in different ways. The responses analyzed showed that white participants attributed the behavior as due to personal factors, such as violent dispositions when the perpetrator was racialized black; and more frequently interpreted the same behavior as due to situational factors when the perpetrator was racialized white (Duncan 1976). In these studies, then, in-group favoritism manifests itself in the disposition to reach for explanatory resources that are favorable to members of our in-group.

A notable feature of this research paradigm, however, is that whilst in some of the studies above the in-group identities tracked social identities that are consequential in structuring social arrangements, and indeed are sometimes identities with which group members positively identify, the in-group bias need not depend on such social identities. In studies that asked participants to allocate rewards to in-group and out-group members, the group identities activated were entirely arbitrary, and artificially constructed by the experimental context: e.g. the tendency to over-estimate or under-estimate the number of dots presented in a prior task (Taifel & Turner 1986). Yet, even still, being a member of one or other of these groups sufficed to activate in-group bias: members of each group typically favored the allocation of resources to in-group members, and sought to maximize the difference in resources allocated to in-group and out-group members.[3] Other studies focused on arbitrarily constructed group statuses that tracked the team color to which participants belonged (being assigned to the 'green' group, or the 'red' group, say) (Dawes et al 1990). The key point to note, then, is that whilst the tendency to form such in-group preferences seems a pervasive feature of our psychology, the in-group/out-group statuses that this form of favoritism may track seems to be highly malleable. That is to say: who counts as in one's in-group may be variable, with

new (sometimes arbitrary) group statuses being formed depending on social contextual features or prompts.

Note that analyses of discrimination have characterized group discrimination such that these sorts of arbitrary in-group preferences would not meet the criteria for social salience. For example, the notion of social salience incorporated by Lippert-Rasmussen has it that:

> A group is socially salient if perceived membership of it is important to the structure of social interactions across a wide range of social contexts (2013: 30).[4]

This analysis of social salience permits us to capture the notion of discrimination as it is used to explain what is wrong with certain kinds of treatment that disadvantage women, racial minorities, certain religious identities, or the elderly, for example. These kinds of social identities clearly structure social interactions and social hierarchies in a broad range of contexts. Note, though, that this analysis will not allow the sort of differential treatment manifested in some of the studies on in-group favoritism to be diagnosed as discrimination. Certainly the disposition to over- or under-estimate dots is not socially salient in this sense: it does not structure interactions across a wide range of social contexts. Yet it is precisely because the experimenters brought this status to salience - as structuring *that* social context - that we saw patterns of differential treatment emerge within the laboratory studies.

Moreover, the experimentally found tendency for in-group bias has been identified by philosophers as one of the mechanisms by which structures of hierarchy and inequality might be perpetuated. Consider Elizabeth Anderson's remarks, in the context of epistemic discrimination, whereby some agents are judged to be less credible than others. She observes that if in-group biases in this context line up with social identities that are already the basis of systematic disadvantages, then we should see this sort of epistemic discrimination as structural (Anderson, 2012: 170). For example, if in-group bias about credibility judgments tracks race, then this will reinforce epistemic and other dimensions of disadvantage that black and minority ethnicity individuals may already

face. Compare this with a context in which in-group bias about credibility judgments track whether one is an over- or under-estimator of dots. This will not track any pre-existing systemic disadvantage. Anderson's remarks alert us to the distinction between the identities that populate an in-group/out-group bias, and the social salience of those identities in structuring or perpetuating wider patterns of social disadvantage.

The point of this observation is to show that analyses of group discrimination that aim to focus on the socially salient identity have a choice to make about the scope of the phenomena to be captured. The choice made will have implications for how social salience is characterized. For example, on the one hand, we may seek to restrict discrimination to those cases that meet the social salience criteria as characterized above. One motivation for doing so is to capture ordinary usage of the notion of discrimination. This would be to maintain that differential treatment on the basis of these arbitrarily drawn and artificially constructed in-group identities is not discrimination. On this construal, dispositions to favor in-group members may be present across a range of differential treatments, but only some of those instances constitute group discrimination proper.

On the other hand, we might seek to treat the in-group biases revealed in various experimental contexts as continuous with the notion of discrimination as captured by the analysis offered in section 1. In that case, we could refine our analysis of social salience, indexing it to a context, such that:

> A group is socially salient, in context C, if perceived membership of it is important to the structure of social interactions in that context, C.

On this construction, in-lab differential treatment such as that described above will count as discrimination, though this manifestly will not transmit to other contexts, in which other identities are salient. What may be socially salient for the purposes of an individual's differential treatment may not coincide with what is socially salient in structuring broader patterns of social disadvantage.

What attention to the social psychological research helps us to see, then, is that the original analysis of group discrimination builds in certain strong conditions for social salience, present only

in some of the cases of differential treatment; albeit those which garner great concern and feature in ordinary usage of the notion of discrimination. But this is just one option: a context sensitive account of salience may be preferred, if one seeks to capture all cases of differential treatment based on in-group bias, and identify the continuity, in terms of operative cognitive mechanism, between the 'in lab' and 'real world' phenomena.

**4. Discrimination involving implicit bias**

In the past couple of decades, social psychologists have focused on the measurement of 'implicit bias', or what is sometimes called 'aversive racism' or unconscious or automatic discrimination. In this section I set out some of the research on implicit bias, and show how our understanding of discrimination – in particular, condition iii of the analysis from earlier – requires revision.

Implicit biases are automatic patterns of thought or feeling.[5] Notably, these patterns may not be transparent to us, and it may be difficult for us to detect and control when they are operating in our minds or influencing our judgments and behaviors. Those which have occupied the attention of philosophers typically involve attaching some negative property or stereotypic trait to membership of a social group. For example, indirect measures have indicated that men are more strongly associated with leadership and women more strongly associated with nurturing and supporting roles (AAUW report 2016).[6] Other studies have found that white people are more strongly associated with intellectual features and black people more strongly associated with physical traits (Amodio & Devine 2006). Other measures have focused on the problematic biases we might have towards people on the basis of religion, age, disability, sexuality (see Jost et al 2009 for an overview of research on implicit biases). Crucially, individuals who sincerely profess egalitarian values and beliefs may display negative associations on these indirect measures. Whilst the mere presence of such biases is troubling, their potential to influence action is particularly relevant to concerns about discrimination. Various studies indicate that behavior may be influenced by implicit bias.[7] Consider first a study on hiring selections by Dovidio & Gaertner (2000). These studies indicated that

individuals were more likely to make hiring recommendations in the case of moderately good white candidates than moderately good black candidates. (In cases where the candidates were obviously stellar or unqualified, the decisions tracked quality of application rather than race). The moderately good candidates had exactly the same credentials, but differed only with respect to the race indicated on their application materials (cued with racially stereotyped names). Accordingly, the differential recommendations appear to be race-based discrimination, and involve a pattern of treatment that would - does - significantly disadvantage black individuals. Yet individuals who made such discriminatory recommendations may not have been aware that they were discriminating, and indeed may avow anti-racist norms of fair treatment when asked to report on their explicit or endorsed beliefs and values.

Second, consider the now well known 'shooter bias' tests (Glaser et al 2006, Correll et al 2007). In these studies, individuals participate in a simulation task whereby they are shown images of males in different postures holding a range of objects. In some cases the object is a weapon (gun); in other cases, it is not (mobile phone). Some of the images are of black males, others of white males. Participants in the simulation are instructed to 'shoot' at only those individuals who are armed. Results indicate that the participants more likely (and faster) to shoot black males with weapons than white males with weapons, and more likely to make the error of shooting an unarmed black male. It is clear how such biases and discriminatory behavior could have gravely disadvantaging effects for black individuals.

Cases such as these – where individuals are discriminating based on race – may not fit our usual understanding of discrimination: the discriminatory treatment may be unintentional. Indeed, that she is engaging in unequal treatment may be something of which the discriminator is unaware; and, were she cognizant of it, would disavow. Nonetheless, I take it as uncontroversial that such patterns of race-based differential treatment, that are obviously disadvantageous, do count as discrimination, and should be captured by an analysis of the notion.[8] However, the conditions we

set out above need refinement if they are able to do so. Let us consider the first three conditions for group discrimination set out in section 2:

> An agent, X, discriminates against someone, Y, in relation to another, Z, by Φ-ing (e.g. hiring Z rather than Y) if and only if:
>
> i) There is a property, P, such that Y has P or X believes that Y has P, and Z does not have P or X believes that Z does not have P,
>
> ii) X treats Y worse than he treats or would treat Z by Φ-ing, and
>
> iii) It is because (X believes that) Y has P and (X believes that) Z does not have P that X treats Y worse than Z by Φ-ing

Take the shooter bias studies. The discriminator (X) is in this case one of the participants in the study. The property in question, P, is the race of the individual in the scenario - P denotes the fact that an individual (Y) is racialized black (or believed to be black). In shooting an unarmed black male in the study images, X treats that individual worse than he treats other individuals in the study images who are not black, with potentially fatal consequences.[9] Insofar as the race of the individuals in the images appears to be the only factor that could explain the differential responses, it appears that it is this perceived property (P) (possessed by Y but not by Z) which explains why X treats Y in that way. In our current context, race is a socially salient group, clearly; and the kind of action at issue - shooting - is such that it is of a relevant type of act of which there are many of this type which makes black individuals worse off (facing increased risk of violence in their daily lives).[10] The point I want to press is that some refinement is needed to the analysis of direct discrimination if it is to be able to adequately capture this sort of discriminatory behavior. The adequacy of the analysis turns on the way in which the third condition is understood. Recall this condition requires that:

iii) It is because (X believes that) Y has P and (X believes that) Z does not have P that X treats Y worse than Z by Φ-ing

We should want to say that it is because (the discriminator – the participant in the test – believes that) the individual is black - and (the discriminator believes that) other individuals in the simulation images are white - that the discriminator treats the black individuals worse (shooting more readily, including at unarmed individuals) than the white individuals.

However, we need to see further what it means to say that someone acted *because* the individual has (or is believed to have) the relevant property (in this case, being racialied black). Lippert-Rasmussen suggests that, in the case of direct discrimination:

> X treats Y worse than Z by Φ-ing because (X believes that) Y has P and (X believes that) Z does not have P if, and only if, the thought that Y, and not Z, has P is part of X's direct, motivating reason for Φ-ing.[11]

This is intended to capture the relation that holds between (perceived or imagined) membership of a socially salient group and disadvantageous treatment as we find it in cases of direct discrimination.

This characterization seems to serve well in cases of discrimination in which explicit bias has a role; that is, in which someone's perceived group membership considered by the agent to be a reason for treating them disadvantageously. Such cases of explicit bias include, for example, if an individual is believed to be less suitable for a job - represented as incompetent - *because* of beliefs about race and competence, or judged less likely to be an effective leader - represented as indecisive - *because* of beliefs about gender and leadership.

Cases of implicit bias discrimination seem not to involve such clear beliefs or judgments that are taken to justify differential treatment. Yet we should expect discrimination due to implicit bias to be captured by an analysis of direct discrimination:[12] it is a matter of an individual treating

another individual disadvantageously on account of her race (or gender, or age etc.). But in fact, whether the analysis offered can capture these cases of discrimination will depend on exactly how condition iii is glossed, and in particular on how we understand the notion of a 'motivating reason' for action. On one broad rendering of this notion, we might follow a Williams-inspired understanding of motivating reason, which is to say that an agent has a motivating reason for some action (φ-ing) when she 'has some motive which will be served or furthered by [her] φ-ing" (1981: 101). On this reading of what it is to have a motivating reason, we commit only to the idea that the agent has some motive that causes her to act. She need not be aware of this motive, or the role it plays in her action; indeed, she may disavow it. Thus construed, the shooter bias case could be understood such that the fact that the individual is racialized black *does* serve as a motivating reason for the participant's differential (worse) treatment – even though the participant is unaware of this motive and its role in her action. Note we must also suppose, on this reading, that 'the thought that' the individual is racialized black must also be construed permissively: as including unconscious thoughts, as well as explicitly entertained reflections.

Whilst this interpretation of Lippert-Rasmussen's understanding - of what it is to treat someone differently 'because' of her perceived group membership - can accommodate cases of discrimination due to implicit bias, there is perhaps something inelegant or unsatisfying about it. In particular, it seems to require some specification of the motive that is furthered by the agent's discriminatory action. There may be answers to be given here – vested and unarticulated or disavowed interests in preserving racial hierarchy (for racialized white participants, at least); or perhaps more simply an interest in efficient cognizing, which promulgates dependency on such fast associative thought. But we will need an answer (or answers) that cover all of the cases in which discrimination due to implicit bias is found, and that is not an ad hoc postulate to ensure the agent comes out as having the requisite motivating reason(s). Perhaps there is work to do in pinning this part of the analysis down so that apprehension of the individuals group membership is the cause - the motivating reason - for the differential treatment.

Note that there are some interpretations of 'motivating reasons' that will not be compatible with Lippert-Rasmussen's analysis. For example, on another rendering of the notion, a motivating reason is 'a reason for which someone does something, a reason that, in the agent's eyes, counts in favor of her acting in a certain way' (Alvarez, 2016).[13] But note that on this latter reading, in many cases of discrimination due to implicit bias, 'the thought that Y has P' (e.g. the individual is black) cannot properly be construed as the agent's 'motivating reason' for action. This is because in many cases, the agent who is influenced by implicit bias disavows the associations (or associated propositions) that are implicated in the discriminatory actions. For example, in the shooter bias case, individuals who display this discriminatory pattern of shooting may nonetheless deny that 'the thought that the individual is black' counts in favor of her acting a certain way. On explicit self-reports, the agent may strongly resist subscribing to propositions such as 'black people are more likely to carry weapons', and repudiate any such racist stereotypes. That the individual is black does not provide the agent with a motivating reason for differential treatment, then. Of course, we might uncharitably take such self-reports as disingenuous and insincere. But a further reason for maintaining that this cannot be posited as the agent's motivating reason is that, in the case of implicit bias, individuals may not intend to discriminate, and indeed may not be aware that they have been implicated in differential treatment.

The upshot of this is that the gloss on motivating reasons needs to be spelled out in more detail. If one spells it out in terms of the broader reading – the motive for or cause of the agent's action – then implicit bias discrimination may be captured by condition iii. However, if for independent reasons one endorses the narrower construal of motivating reasons, a different articulation of condition iii will be needed in order to capture the phenomena of discrimination that results from implicit bias. In these cases, what is needed is a more direct connection between the socially salient group membership and the differential treatment; one that is not mediated by the agent's thoughts about her reasons for action. Various options are available. For example, we might say that:

> X treats Y worse than Z by Φ-ing because (X believes that) Y has P and (X believes that) Z does not have P if, and only if, 1) the thought that Y, and not Z, has P is part of X's direct, motivating reason for Φ-ing, or **2) X's categorisation of Y, and not Z, as P influences her Φ-ing.**

Whilst 1) continues to deal with the cases where discrimination is intentional and based on explicit prejudice, this additional clause, 2), should capture cases where the agent's automatic processing of social identity, and automatically activated stereotypes or affective states, influence the agent's behavior without her awareness, and without her intention. The mere (and perhaps non conscious or subpersonal) categorization of the individual as belonging to a socially salient group may suffice to activate the automatic processes that influence the agent's behavior.

This might seem excessively complicated. An alternative fix would be to appeal to a contrast class that is on the table with talk of motivating reasons: namely, explanatory reasons. An explanatory reason is one that *explains* an agent's action, and it certainly seems clear that we should want to appeal to social category information (that the individual was racialized black) to explain the agent's differential treatment. This would lead us to the following specification of condition iii:

> X treats Y worse than Z by Φ-ing because (X believes that) Y has P and (X believes that) Z does not have P if, and only if, Y having (or being believed to have) P, and not Z, **is the explanatory reason for X's Φ-ing**.

This is certainly a more elegant fix. However, one reason to worry about this way of capturing implicit bias discrimination is that we may then lose the distinction between direct and indirect discrimination. This is because in cases of indirect discrimination, the fact (for example) the agent is racialized black will also *explain* the differential treatment, but indirectly, via the proxy

consideration that ends up yielding racial discrimination. To accommodate this, one could restrict explanatory reasons to those which are psychological: namely, those which feature in the agent's psychology, rather than those which have *any* explanatory role. Since (I assume) we should want to maintain the distinction between direct and indirect discrimination, and see discrimination due to implicit bias falling on the direct discrimination side of the distinction, this fix will have to be understood in this way to do the work required.

There may be further reasons to object to these modified conditions, or to favor one over the other. The main concern here, though, is to demonstrate that attention to the research in social psychology has import for analyses of discrimination. If these analyses seek to capture such phenomena as discrimination due to implicit bias, then some revisions - of the sort proposed above - may be in order.

**4. Stereotype threat**

A third domain of research in social psychology that is worth turning our attention to concerns the phenomena known as 'stereotype threat'. As detailed by Steele (1997, 2010), 'stereotype threat' refers to the cases in which the perceived threat of confirming a stereotype has a deleterious effect on an individual's performance on a task.

For example, Steele and Aronson (1995) asked black and white college students to take a study comprised of questions from the GRE (Graduate Record Examination). Some of the participants were told the test was diagnostic of intellectual ability, whilst others were explicitly told it was not diagnostic of abilities. The diagnostic condition activated stereotypes about the lower intellectual abilities of black people. In this condition, black students underperformed compared to the white students; no such effect was found in the non-diagnostic condition. Shih et al (1999) found that stereotypes about gender had a similar effect: when gender - and the perceived threat of the stereotype that women are poor at math - was made salient, women students' performances suffered. Interestingly, this effect disappeared when another aspect of the women's identities were

made salient, namely, their Asian-American identity, which is stereotyped as skilled at quantitative reasoning.

These studies, and others, are taken to support a variety of hypotheses about the mechanism by which perceived stereotype hinders performance. Candidate explanations contend that the activation of a stereotype, and an individuals' awareness of this stereotype - even if they do not endorse it - may provoke anxiety. Alternatively, such awareness may decrease motivation on the task; or cause subjects to unintentionally imitate and express the stereotype in their engagement with the task; or simply create self-doubt which distracts from the task in hand.[14]

The phenomenon of stereotype threat raises some interesting questions for proposed analyses of discrimination. In particular, we may press for greater precision with respect to our understanding of what it is to 'treat' an individual X worse than Y - the framing that appears in condition ii) of the analysis above (section 1). Does the sort of behavior that is described as 'activating a stereotype', in the context of these studies, constitute a form of differential (worse) treatment? In some cases this 'activation' amounted to no more than merely mentioning gender, or the mere mention of some feature of the task associated with a stereotype - for example, whether a test is diagnostic of intellectual abilities. Steele and Aronson (1995) simply told their participants (black and white) in the diagnostic condition that:

> 'because we want an accurate measure of your ability in these domains, we want to ask you to try as hard as you can to perform well on these tasks. At the end of this study, we can give you feedback which may be helpful by pointing out your strengths and weaknesses' (803).

It is rather surprising to consider this prompt as a form of differential and disadvantageous treatment that partially constitutes discrimination. This will be especially so in cases in which the individual is not at all aware that such words activate stereotypes that can have a deleterious effect

on performance. Nonetheless, it is true that this mere mention did, under some description, involve treating individuals involved in the study differently and disadvantageously. For black students, this constituted *activating a negative stereotype* - since this prompt sufficed to activate negative stereotypes concerning black people and intellectual abilities, and thereby caused (via one or other of the candidate mechanisms mentioned above) lower levels of performance. For the white students who were given this very same prompt, no negative stereotype was activated.

Should we capture such behaviors, which (perhaps unknowingly) activate stereotypes, as directly discriminatory *treatment*? How one might unpack the notion of 'treatment' will depend on whether one is inclined to subsume the phenomenon of stereotype threat under the rubric of direct - rather than indirect - discrimination. It seems unclear to me what intuition or ordinary usage supports, and broader theoretical concerns might guide one's choice here. However one chooses, there is some work to do. If one is so inclined to see stereotype threat as a form of direct discrimination, then a notion of 'treatment' that can handle the various and subtle ways of *activating stereotypes* will be needed. On the other hand, one may not want to extend the analysis of direct discrimination to include stereotype threat. On this line of thought, we could leave such cases to be captured by the analysis of indirect discrimination.[15] Then an understanding of 'treatment' should be developed, as part of the analysis of direct discrimination that excludes such activation. This might be a delicate matter, however, since manifestly, we have reason to maintain that treating (in the standard common-sense way) according to stereotypes is a form of discrimination (cf. Lippert-Rasmussen at p.21). Yet, we have seen that direct discrimination need not be intentional (section 4 above; see also Lippert-Rasmussen fn 70). So, circumscribing the notion of 'treating' in a way that maps these complex contours, and allows us to allocate stereotype threat to the desired side of the direct/indirect distinction, may be a demanding task.

**5. Concluding remarks**

We have seen three domains of research from social psychology, concerning in-group favoritism; implicit bias, and stereotype threat. Each poses distinctive challenges to philosophical analyses of discrimination: the first concerns how we might characterize the 'social salience' of group membership on the basis of which one is disadvantageously treated, and whether this should include all cases of in-group favoritism, or only those that track groups that have widespread social salience. The second concerns how the basis for a discriminator's differential treatment is understood, and whether it need make reference to the explicit beliefs of the discriminator, or include mere automatic categorizations that instigate implicit biases. Finally, we saw that some refinements may be needed to what is included within the scope of differential *treatment*, and whether it includes seemingly innocuous behavior that nonetheless activates stereotypes, and thereby stereotype threat. Whilst these matters are philosophical, and cannot be settled simply by appeal to empirical research, my hope is that these observations provide further motivation for philosophers interested in discrimination to mine the rich seams of research on the social psychology of discrimination.[16]

**Biographical Note**:

Jules Holroyd is a Vice-chancellor's Fellow in Philosophy at the University of Sheffield. She has published on the philosophy and psychology of implicit cognition, and is principle investigator on the Leverhulme Trust-funded research project on Bias and Blame.

**Bibliography**

AAUW report 2016 http://www.aauw.org/resource/barriers-and-bias/

Al Ramiah, A. Hewstone, M., Dovidio J.F. Penner, L.A. 2010 The Social Psychology of Discrimination: Theory, Measurement and Consequences in L. Bond, F. McGinnity & H. Russell (eds.) *Irish and International Approaches to Measuring Discrimination* Liffey Press, pp. 84-112.

Alvarez, M. 2016 ‘Reasons for Action: Justification, Motivation, Explanation’, *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/win2016/entries/reasons-just-vs-expl/>.

Amodio, D. M., & Devine, P. G. 2006. ‘Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior'. *Journal of Personality and Social Psychology*, Vol. 91, No. 4, 652–661

Anderson, E. 2012 ‘Epistemic Justice as a Virtue of Social Institutions’ *Social Epistemology* DOI: 10.1080/02691728.2011.652211

Baber, H.E. 2001. ‘Gender Conscious’, *Journal of Applied Philosophy* 18.1

Brewer, M. 1999. ‘The psychology of prejudice. Ingroup love or outgroup hate?’ *Journal of Social Issues* 55 (3): 429–44.

Blank, R.M., Dabady, M. and Citro, C.F. (eds.) 2004. *Measuring Racial Discrimination* Washington, DC: The National Academies Press.

Correll, J., Park, B., Judd, C.M., Wittenbrink, B., Sadler, M.S. and Keesee, T. 2007. 'Across the thin blue line: Police officers and racial bias in the decision to shoot'. *Journal of Personality and Social Psychology, 92*, 1006- 1023.

Dancy, J., 2000. *Practical Reality*, New York: Oxford University Press.

Dovidio, J.F. and Gaertner, S.L. 2004. 'Aversive racism'. In M. Zanna (ed.), *Advances in experimental social psychology* (pp. 1- 52). San Diego, CA: Academic Press.

Dawes, R., van de Kragt, A. and Orbell, J. 1990. ‘Cooperation for the benefit of us—not me, or my conscience’. In *Beyond self-interest*, edited by Jane Mansbridge, pp. 97– 110. Chicago: University of Chicago Press.

Duncan, B. L. 1976. ‘Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks’, *Journal* of *Personality and Social Psychology,* 34:590-598.

Glaser, J. Knowles, E. 2008 ‘Implicit Motivation to Control Prejudice’ *Journal of Experimental Social Psychology*, Vol. 44, pp. 164-172

Hewstone, M. and Ward, C. 1985. ‘Ethnocentrism and causal attribution in Southeast Asia’, *Journal of Personality and Social Psychology*, 48: 614-623,

Hewstone, M. 1990 ‘The ‘ultimate attribution error‘? A review of the literature on intergroup causal attribution’ *European Journal of Social Psychology*, Vol. 20,311-33.5

Holroyd, J. 2016 ‘What do we Want from a Model of Implicit Cognition?’ *Proceedings of the Aristotelian Society* 116(2) pp.153-179

LeBron, C. 2016 ‘I’m Black. Does America Have a Plan for My Life?’ *The Stone* Sept 26, http://www.nytimes.com/2016/09/26/opinion/im-black-does-america-have-a-plan-for-my-life.html

Levy, N. 2014, ‘Neither fish nor fowl: Implicit attitudes as patchy endorsements’, *Noûs*. doi:10.1111/nous.12074

Lippert-Rasmussen, K. 2013 *Born Free and Equal: A Philosophical Analysis of Discrimination*, Oxford University Press.

Machery, E. 2016 De-Freuding Implicit Attitudes, in Brownstein, M. & Saul, J. (eds.) Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology.

Mallon, R. 2016 Stereotype Threat and Persons, in Brownstein, M. & Saul, J. (eds.) Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology, pp.131-154

Mandlebaum, E. 2015 “Attitude Inference and Association: On the Propositional Structure of Implicit Bias” *Nous* 10.1111/nous.12089

Oswald, F.L., Mitchel, G. Blanton, H. Jaccard, J. Tetlock, PE., 2013 “Predicting ethnic and Racial Discrimination: A meta-analysis of IAT criterion studies” *Journal of Personality and Social Psychology* 105(2) 171-192.

Rosenberg, S. W. and Wolfsfeld, G. 1977. ‘International conflict and the problem of attribution’, Journal of Conflict Resolution, 21: 75-103.

Shih, M., Pittinsky, T. L., et al. 1999. “Stereotype susceptibility: Identity salience and shifts in quantitative performance.” Psychological Science 10: 80–83.

Steele, C. 2010. Whistling Vivaldi: And Other Clues to how Stereotypes Affect Us. New York, NY: W. W. Norton and Company.

Steele, C.M. 1997. 'A threat in the air: How stereotypes shape intellectual identity and performance'. *American Psychologist, 52*, 613- 629.

Steele, C.M., and Aronson, J. 1995. 'Stereotype threat and the intellectual test performance of African Americans'. *Journal of Personality and Social Psychology, 69*, 797–811.

Stephan, W. G. 1977. ‘Stereotyping: Role of ingroup-outgroup differences in causal attribution of behaviour’, Journal of Social Psychology, 101: 255-266.

Taylor, D. M. and Jaggi, V. 1974. ‘Ethnocentrism and causal attribution in a South Indian context’,*JournalofCross-CulturalPsychology,*5 162-171.

Tajfel, H. and Turner, J. C. 1979. ‘An integrative theory of intergroup conflict’. In: Austin, W. G. and Worchel, S. (Eds) The Social Psychology of Intergroup Relations, BrookdCole, Monterey, CA.

Valian, V. 1999 Why So Slow? The Advancement of Women, MIT Press.

Williams, B. 1981 *Moral Luck*, Cambridge: Cambridge University Press.

Winkler, J. D. and Taylor, S. E. (1979). ‘Preference, expectations, and attributional bias: Two field studies’, *Journal of Applied Social Psychology,* 9 183-197.

[1] Note that whilst these conditions capture the sorts of treatment that we find morally objectionable, no moral evaluation is built into the conditions themselves: it is thus a contingent empirical fact that these conditions pick out treatment that is often morally objectionable - often in a way that makes it wrong - when they do. (Lippert-Rasmussen's later analysis of the wrong as to do with the harm that discrimination imposes fills out that part of the account - see chapter 6 of his 2013).

[2] However, for a more complex pattern of results, see Hewstone & Ward (1985), who focused on Malay/Chinese identities; and Stephan (1977) which examined attributions of anglo- Chicano- and black American high school students. All these studies are discussed in the overview by Hewstone 1990.

[3] Also worth noting is that in-group favoritism need not coincide with antipathy towards an out-group (Brewer 1999; see also Ramiah el at 2010 p.88).

[4] Compare also Baber's notion of social salience, which focuses on the extent to which a property is taken to 'predict and explain beliefs, character traits, tastes or other socially significant psychological characteristics' (2001: p.53). This characterization will also exclude from salient those arbitrary and artificial in-group statuses, insofar as whether one is (e.g.) an under or over-estimator is clearly not taken to predict or explain much at all.

[5] Contention abounds as to how to characterize these automatic thought processes – whether they involve associative mental states, states with propositional content, affective or representational content, and whether they are mental states at all. For discussion see Mandelbaum 2015, Levy 2015, Machery 2016, Holroyd 2016.

[6] Downloadable here: http://www.aauw.org/resource/barriers-and-bias/

[7] There is contention over the predictive validity of implicit biases: see discussion in Machery 2016, and Oswald et al 2013. The important point to take away from the research, it seems, is that biases may lead to discriminatory behavior, though it is difficult to determine when this is so. Given this, we cannot be confident that our behavior is not influenced in these discriminatory ways. For debate about these concerns regarding predictive validity, see the roundtable discussion at The Brains Blog: http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

[8] See discussion from Christopher LeBron (2016) of the wide reaching impact of shootings of black males, by police, on the lives of all racialized as black. Of course, we need not believe all such cases are due to implicit bias.

[9] This is slightly complicated by the fact that we are talking about a fictional individual, portrayed in the study images. However, I do not take this complication to affect the substantive point at issue here: that such discriminatory behavior cannot be captured by the conditions as stated.

[10] Of course, the study is alarming insofar as it indicates that such behavioral propensities might be found outside of laboratory simulations. As such, we might cautiously say that such studies are evidence of discrimination outside the lab, rather than acts of discrimination themselves. But my point is not that the laboratory setting cannot be accommodated by this set of conditions. The point is that the analysis cannot as it stands capture differential treatment that results from such biases.

[11] Note that Lippert-Rasmussen appears to endorse this condition as an analysis of direct discrimination (37), but rejects it as an analysis of discrimination tout court, since indirect discrimination cannot be accommodated by this analysis. The final and full articulation of this condition for group discrimination is disjunctive, whereby: 'X treats Y worse than Z by Φ-ing because (X believes that) Y has P and (X believes that) Z does not have P if, and only if, (i) the thought that Y, and not Z, has P is part of X's direct, motivating reason for Φ-ing, or (ii) the fact that Y, and not Z, has P causally explains X's Φ-ing and this in turn is causally explained by the fact that people with P are often treated worse than those without P in the sense given by (i)' (38). The second disjunct here is supposed to capture the phenomena of indirect discrimination. However, I take it for our purposes, our attention should be on the first disjunct (i) since discrimination due to implicit bias should be captured by an analysis of direct, rather than indirect discrimination.

[12] Note that in some cases, implicit bias is revealed across groups of individuals. For example, consider studies, such as the CV studies mentioned above, where the structure is a 'between participants' study, such that no participant makes judgements about or interacts with both black and white individuals, and differential treatment emerges in patterns of behavior across participants - some of whom interact with or make evaluations about black individuals, others of whom interact with or make evaluations of white individuals. In such cases, our understanding of discrimination due to implicit bias may pose additional difficulties for this model of direct discrimination. The

analysis will perhaps only hold on the assumption that, in these 'between participants' studies, each individual would have judged differently were the race of the applicant different. My remarks in the text above apply to cases where implicit bias is revealed by differential treatment, by one person, of individuals differentially racialized.

[13] See also Dancy's view of motivating reasons, whereby for something to be a motivating reason for the agent, she has to take that consideration to be a normative reason for acting (Dancy 2000). This construal of motivating reasons similarly seems to require that the consideration 'in the agent's eyes' is a reason for – a consideration in favor of - action.

[14] For an overview of a range of studies on stereotype threat, and an evaluation of the competing hypotheses, see Mallon 2016.

[15] For the statement of the analysis of indirect discrimination, see footnote 11. The key difference is the focus on the causal role of social category information, rather than the role it plays in the psychological dispositions of the individual who disadvantageously treats another.

[16] This research was completed with the support of the Leverhulme Trust project grant on Bias and Blame (RPG-2013-326).