Oxford Journal of Legal Studies, Vol. 35, No. 1 (2015), pp. 121–152
doi:10.1093/ojls/gqu025
Published Advance Access November 12, 2014

Taming the Biased Black Box? On the Potential Role of Behavioural Realism in Anti-Discrimination Policy

Ana Carolina Alfinito Vieira* and Alex Graser**

Abstract—Anti-discrimination laws have long been established in many legal systems, and the relevant body of rules has constantly grown. But findings from social psychology research suggest that these policies are based on unrealistic premises and are therefore bound to remain unsuccessful in many instances. While legal scholarship has begun to reflect upon these insights and to discuss a number of individual policy responses, this essay seeks to provide a more comprehensive framework within which the implications of implicit social cognition for anti-discrimination policies can be understood, and to map out the range of reform options for anti-discrimination policy.

Keywords: anti-discrimination policy, equality-oriented policies, implicit social cognition, implicit bias, affirmative action

1. *Introduction*

Most polities would nowadays consider it a primary goal to ensure equal opportunity for all their members—or even for everybody within their jurisdiction. Governments adopt a multitude of policies to that end, and as they realise that this goal can be compromised by innumerable individual decisions, perpetually, day to day, many of these policies are directed at controlling such individual decisions. Anti-discrimination laws[1] are an example of those kinds of policies—not the only one, to be sure, but the primary one.

* Doctoral Fellow at the Max-Planck-Institute for the Study of Societies, Cologne. Email: av@mpifg.de.
** Professor of Public Law and Policy at the University of Regensburg and Fellow of the Hertie School of Governance, Berlin. Email: Alexander.Graser@ur.de. The present research was supported by grants from the Deutsche Forschungsgemeinschaft and MaxnetAging, a project of the Max Planck Society.

[1] For a concise and comprehensive treatment of these policies from a comparative perspective, see S Fredman, *Discrimination Law* (2nd edn, OUP 2011).

© The Author 2014. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

MPIfG Journal Article
Ana Carolina Alfinito Vieira, Alexander Graser: Taming the Biased Black Box? On the Potential Role of Behavioural Realism in Anti-discrimination Policy. In: Oxford Journal of Legal Studies 35(1), 121-152 (2015). Oxford University Press
The original publication is available at the publisher's web site: http://dx.doi.org/10.1093/ojls/gqu025

The MPIfG Journal Articles series features articles by MPIfG researchers and visiting scholars published in peer-reviewed journals.
Max Planck Institute for the Study of Societies (MPIfG) Cologne | www.mpifg.de

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

These policies, however, have often proved ineffective. The reasons for such failure are manifold. Enforcement is difficult and may come at a high price, especially as individual actors will often have strong incentives to circumvent such interference with their decision-taking. Regulatory efforts have long been concerned with these challenges. The level of differentiation that anti-discrimination laws have achieved in many jurisdictions testifies to these efforts. But the outcome remains less than satisfactory. Disadvantage persists. In many polities, a person's skin colour still determines to a significant extent what opportunities she may enjoy, and the same can be said about her gender, age, and many other personal features.

Hence, there is ample reason to reconsider the traditional strategies in anti-discrimination policy. This is all the more warranted as research in social psychology has increasingly called into question the basic assumptions underlying such policies. The message, in a nutshell, is that the addressees of anti-discrimination laws may be incapable of complying with these commands even if they are willing to do so. The reason is that perceptions, attitudes, beliefs, and actions may be biased without the respective person being aware of this, let alone able to correct it. In such cases, it will obviously not work simply to require an individual not to act in a discriminatory fashion, which, however, is exactly what anti-discrimination laws typically do. So this may indeed present a fundamental challenge.

Such findings reveal the discrepancies that exist between, on the one hand, the behavioural assumptions of traditional anti-discrimination law and, on the other, what the empirical social sciences tell us about how people perceive, evaluate, decide, and ultimately act. But how should policy makers react to this discrepancy? In adopting a behavioural–realist approach,[2] this article proceeds from the assumption that law and public policy need to be grounded in scientifically sound accounts of real world phenomena. This does not imply any *a priori* policy preferences, nor does it presume that science could dictate what is desirable or legitimate.[3] But whenever law and policy rest on inaccurate conceptions of real-world conditions, behavioural realism in the above sense would call for their reconsideration and possibly also for their adjustment. This is particularly true when such law and policy fail to fully achieve their proclaimed goals, which, we submit, is presently the case with traditional anti-discrimination law.

[2] For in-depth accounts of the philosophical and methodological premises of behavioural realism, see J Kang and M Banaji, 'Fair Measures: A Behavioral Realist Revision of Affirmative Action' (2006) 94 Cal L Rev 1063; and L Hamilton Krieger and ST Fiske, 'Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment' (2006) 94 Cal L Rev 997.

[3] As noted by Krieger and Fiske (n 2) 1061, 'Behavioral realism does not attempt to introduce social science into normative legal reasoning. Rather, it seeks to extract from normative legal reasoning the intuitive social science already there and to subject it to empirical scrutiny. Understood as a normative theory of adjudication, behavioral realism seeks to hold judges accountable for their rhetorical use of empirical propositions that have in fact been invalidated by advances in the empirical sciences.'

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

This article is part of a growing body of literature that discusses the societal and legal implications of the research programme on implicit bias.[4] We seek to contribute to this literature by spelling out the challenges that the findings of implicit social cognition pose to traditional anti-discrimination law, and by systematically mapping out and discussing the range of policy interventions that are available to counter and circumvent (the effects of) bias in individual decisions. We thus hope to provide a scientifically informed update of the basket of anti-discrimination policies.[5]

First, we shall provide a rough sketch of the relative importance of anti-discrimination laws within the broader field of equality-oriented policies (section 2). Next, we will review the relevant research in the field of implicit social cognition (section 3). Finally, we shall discuss the implications for anti-discrimination law and policy (section 4).

2. *Equality-oriented policies and anti-discrimination law—an evolutionary sketch*

Although it is common for modern governments to pursue equality as an important policy goal, polities differ with regard to the kinds and extent of inequalities that they deem acceptable, and, relatedly, also with regard to the reach and targets of the equality-oriented policies that they employ. Despite such variation, it is, however, possible and indeed useful as a background for our argument to highlight some long-term trends in equality-oriented policies.[6]

A. *Three Layers of Equality-Oriented Policies*

First, one may distinguish three major layers of equality-oriented policies. We speak of 'layers' rather than 'waves' or 'stages' because they do not typically replace but rather add to one another.

[4] For reviews of the literature, see AG Greenwald and L Hamilton Krieger, 'Implicit Bias: Scientific Foundations' (2006) 94 Cal L Rev 945; SR Bagenstos, 'Implicit Bias, "Science", and Antidiscrimination Law' (2007) 1 Harv L & Policy Rev 477; J Kang, 'Trojan Horses of Race' (2005) 118 Harv L Rev 1489; and JT Jost and others, 'The Existence of Implicit Bias is beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that no Manager should Ignore' (2009) 29 Organizational Behavior 39.

[5] Throughout this article, the terms 'anti-discrimination policies' and 'anti-discrimination laws' will have different meanings. The former concept is broader. It comprises not only the latter, but also many other policies such as, most notably, affirmative action policies. It is congruent with what we will below call 'third layer equality-oriented policies' (see section 2A(iii)).

[6] For a more extensive account of the concept of equality-oriented policies, see A Graser, 'Políticas Orientadas para a Igualdade: Um Novo Conceito em Política Pública?' (2009) 4 Revista Brasiliera de Direitos Fundamentais e Justiça 13; and reprinted in A Graser, 'Politicas Orientadas para a Igualdade: Um Novo Conceito em Politicas Publicas?' (2011) 12 Revista Latino-Americana de Estudos Constitucionais 12.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

(i) First layer: Equalising the rules of the game

The first layer, in terms of both historical sequence and systematic fundamentality, is general equality of and before the law.[7] It implies an abolition of all status difference and the concurrent operation of a cross-cutting equality principle that is typically situated at some elevated rank within the normative hierarchy. This first layer is often referred to as formal equality as it neither prevents nor even targets material inequalities. Indeed, these may well persist despite the institutionalised recognition of this first layer.

(ii) Second layer: Equalising starting points

This is where the second layer comes into play. It consists of policies that aim not just at equalising the 'rules of the game', but also at levelling the playing field—(re-)distributive policies—which may take the shape of institutional spending, be built into the tax system, or come as individualised social grants. These policies are typical of welfare states, where governments are charged with the task of guaranteeing that all the citizenry have access to the minimal provisions that ensure a good life, including basic levels of education, social assistance and healthcare. Unlike the first layer of formal equality-oriented policies, those of the second would not necessarily nor even typically aim at achieving full equalisation in substantive terms. Generally, their goal is to equalise starting points, and although it might be hard to tell 'starting points' from all other distributive conditions, it is widely thought that too far-reaching substantive equalisation would fail to honour individual achievement and choice of lifestyle.[8] Hence, persisting inequality is not per se a challenge to this second layer of equality-oriented policies. But it is problematic to the extent that it is not attributable to individual choice or achievement.[9] And indeed, many of today's inequalities would hardly seem justifiable by reference to these two factors.[10]

There are many explanations for this. Among the most important is the insight that even if government were able to perfectly design and implement the first two layers of equality-oriented policies, inequality might persist. This is because existing inequality is determined not just by such public action, but continuously reproduced by a multitude of individual daily decisions. Bias with regard to skin colour, gender, age and many more features seems to affect

[7] For a brief historical sketch, see Fredman (n 1) 4–14.

[8] For a concise treatment of the issue, see GA Cohen, *Why Not Socialism?* (Princeton University Press 2009).

[9] See T Parsons, 'Equality and Inequality in Modern Society, or Social Stratification Revisited' (1970) 40 Sociological Inquiry 13 (on the decreasing potential of ascriptive features to justify inequality as a long-term trend in modernity).

[10] For a treatment of individual achievement as insufficient justification for existing wage inequalities, see C Offe, *Industry and Inequality: The Achievement Principle in Work and Social Status* (Hodder & Stoughton Education 1976). For a discussion of the coincidental inequality generated due to birth into a community, see A Shachar, 'Children of a Lesser State: Sustaining Global Inequality Through Citizenship Laws' (2002) Jean Monnet Working Paper 2/03, <http://centers.law.nyu.edu/jeanmonnet/archive/papers/03/030201.pdf> accessed 24 March 2014.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

individual decision-taking in employment relations, rental agreements, medical care, and in many other fields.[11]

(iii) Third layer: Overcoming discrimination

It is this kind of individual bias that the third layer of equality-oriented policies addresses. As with the previous layers, there are many ways in which such individual bias can be targeted, and throughout this article we will keep coming back to the issue of choosing the appropriate policy tools to this end. Also, the reach of the third layer of equality-oriented policies is problematic, at least to the same extent as is true for the second layer of (re-)distributive policies. Even if everybody were to agree in principle that individual bias may be detrimental to social equality, there is disagreement as to whether and to what extent it should be corrected. For any such correction would require an intervention into the sphere that is commonly ascribed to individual autonomy, the sphere, that is, in which the basic assumption is that the individual is free to decide according to her will and that the common good is served best this way. Against this background, interventions need to be justified as exceptions that are both warranted and workable. Third layer equality-oriented policies regularly face objections in this regard.

(iv) Revisiting the layer metaphor

All three layers of equality-oriented policies typically coexist nowadays. Although they have emerged—roughly—in the suggested sequence, this is not a systematic necessity, and none of the layers presupposes a certain level of development of the others. In fact, the pursuit of policies of the respective layers varies across systems, especially with regard to layers two and three.

Also, it should be noted that not all equality-oriented policies can be easily assigned to just one of these layers. Third layer policies may, for example, (be intended to) achieve distributive effects, which could justify classifying them as second layer equality-oriented policies. Anti-discrimination laws pertaining to wage discrimination are a case in point. Moreover, it is not always possible to neatly separate the operation of first layer general equality norms and third layer anti-discrimination laws. This is because their respective scope of application may overlap. General equality clauses will regularly apply to—and at times even explicitly prohibit—instances of discrimination, especially in the public sphere, and these may well include individual decisions. Conversely, anti-discrimination laws may at times extend beyond decisions that are taken by individuals and apply to decisions, policies or rules adopted by firms or in collective agreements.

[11] See JF Dovidio and SL Gaertner, 'Aversive Racism and Selection Decisions: 1989 and 1999' (2000) 11 Psychol Sci 315, for an account of how, despite the widespread institutionalisation of egalitarian values in many modern societies, discrimination continues to exist and affect inequality through aversive racism, a form of pervasive prejudice manifested in ways that are subtle and may even be rationalised.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

It may suffice here to mention these ambiguities. For the purposes of this article, we need not aim for a more exact classification. We should, however, emphasise that our focus is on individual decisions—that is, decisions by individuals—and on the measures that seek to prevent discrimination in such decisions.

B. *The Preference for Traditional Anti-Discrimination Law*

There is a wealth of different equality-oriented policies that fit under our label of third layer policies, and the remainder of this section will be focused on exploring these policies. On the one hand, individual bias may be targeted by measures addressing widespread popular bias in general. Educational programmes for schools are a case in point, as are public awareness campaigns and the like. On the other hand, third layer policies may be addressed specifically at individual decisions and the processes through which they are made. This seems to be the more important of the two basic categories of third layer policies, at least if measured by their contestedness.

This second category comprises various policies. Some of them prescribe a certain outcome for some or all decisions of a certain type. These are typically referred to as affirmative action—*sensu stricto*—such as mandatory quotas for disadvantaged groups in, say, employment relations. Other policies of this category would not refer to the outcome, but only prescribe the procedure in which the respective decisions be taken, such as the mandatory involvement of a gender representative in hiring decisions. Yet another approach is to simply prohibit certain criteria, such as the affected person's gender, skin colour, age etc. This is what anti-discrimination laws typically do.

Within the array of non-general third layer policies, there is a predominance of the latter type, ie, of traditional anti-discrimination laws. The other policies, especially affirmative action rules, might be more effective, but they are often viewed as more problematic. This is to be understood against the background mentioned before, that is, that all third layer interventions infringe upon what is considered the realm of individual autonomy. Traditional anti-discrimination laws would seem to be those that are most respectful of such autonomy. This might explain their apparent predominance.[12]

[12] In addition, affirmative action policies are often viewed as infringing upon the rights not only of the decision-taker, but of other individuals who compete with the beneficiary of these rules. This is an important objection that requires a thorough discussion of aspects largely unrelated to the argument that we present here. For the sake of brevity and focus, we do not deal with this argument further. For an extensive treatment, see A Graser and AC Alfinito Vieira, 'The Case Against the Case Against Affirmative Action' in O Dupper and K Sankaran (eds), *Affirmative Action: A View From The Global South* (forthcoming).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

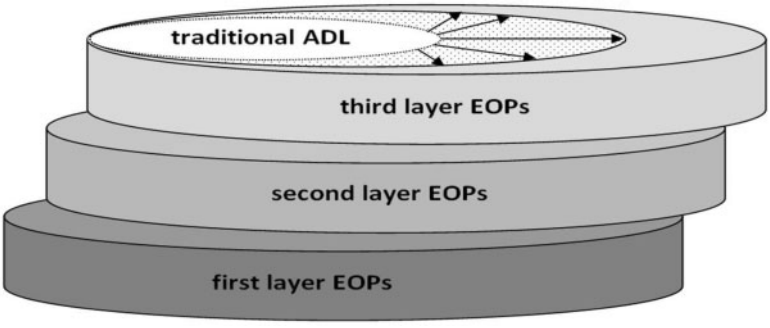

Figure 1. The Relative Weight of Anti-Discrimination Laws (ADL) in Equality-Oriented Policies (EOPs)

C. *The Expansion of Traditional Anti-Discrimination Law*

Over the last decades, traditional anti-discrimination laws have in many polities experienced a considerable expansion.[13] There are two dimensions to this expansion. First, the range of protected groups—or put differently, of criteria that ought not to play a role in individual decisions—has been widened. The sequence of this expansion is highly dependent upon the historical and cultural context. Generally, race and gender were addressed early on, followed by criteria like disability, sexual orientation and age, leading up to, say, weight, as a rather more recent criterion.

Anti-discrimination laws generally do not cover all kinds of interaction that are subject to private law. Intervention in a marriage decision, for example, is hardly conceivable, even if protected traits such as skin colour or age have played a role. Interestingly, the law has traditionally even prescribed that gender has to be considered in these matters, even though this general rule has been eroding in many societies lately. At the other end of the scale, employment relations have for long been made subject to a strict and far-reaching anti-discrimination regime, and an increasing number of other types of private interaction—or sectors of social life—have followed suit. This is the second dimension of the expansion that anti-discrimination laws have undergone since their inception.

We said at the beginning of this article that our argument would focus on anti-discrimination policies. By way of summary to this section, Figure 1 illustrates that these policies are part of the much larger group of

[13] For a general overview on the development of anti-discrimination law see Fredman (n 1) 38–108. For an interesting account of the early, in part even pre-modern roots, of anti-discrimination norms, see T Altwicker, *Menschenrechtlicher Gleichheitsschutz* (Beiträge zum ausländischen öffentlichen Recht und Völkerrecht 223, Springer 2011) 103–7.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

equality-oriented policies, and that they belong to the most recent of the three layers that we have distinguished with regard to these policies. Among the policies of that third layer, traditional anti-discrimination laws are predominant, and moreover, their scope has continuously grown.[14] This may be taken to underscore the considerable importance of the policies that we set out to (re-)assess here. In order to do so, we need to have a look at some insights from social psychology.

3. *Insights from Implicit Social Cognition—Another Evolutionary Sketch*

Since the 1980s, insights from the science of implicit social cognition have led to a profound restructuring of the way human behaviour is understood. Contrasting with 'naïve' psychological conceptions of social behaviour, which assume that individuals are guided solely by their explicit beliefs and willingness to act, research in implicit social cognition has demonstrated that 'actors do not always have conscious, intentional control over processes of social perception, impression formation and judgment that motivate their actions'.[15] These findings have already transformed scientific understanding of human agency. If taken seriously by policy makers, they should also transform the way anti-discrimination policy is framed.

This section will review the findings on implicit mental processes that have major implications for anti-discrimination policy. It will focus on exploring three aspects of intergroup bias: (i) the implicit and automatic dimensions of prejudice and stereotypes; (ii) the measurement and pervasiveness of implicit bias; and finally (iii) the correlation between implicit bias and discriminatory behaviour.

A. *Research on Implicit Stereotyping and Prejudice*

The idea that a large fraction of our mental processes occurs outside the realm of introspective awareness was first introduced over a century ago through psychoanalytical theory and its conceptualisation of the unconscious mind.

[14] For a comparative overview of the prevalence of these (and other) equality-oriented policies in various jurisdictions, see A Graser and D Jackson, *Equality-Oriented Policies – The Concept* (Nomos 2014); and AC Alfinito Vieira and A Graser, *Equality-Oriented Policies in Brazil, India, and South Africa* (Nomos forthcoming).

[15] Greenwald and Hamilton Krieger, 'Implicit Bias: Scientific Foundations' (n 4) 946. For partial reviews of the emergence and consolidation of research in implicit social cognition, see A Greenwald, 'New Look 3: Unconscious Cognition Reclaimed' (1992) 47 Am Psychol 766, on early empirical evidence and measurement of implicit cognition; N Dasgupta, 'Implicit Ingroup Favoritism, Outgroup Favoritism, and their Behavioral Manifestations' (2004) 17 Soc Jus Research 143, on the development of research on implicit ingroup and outgroup preferences and their impacts social interaction; and IV Blair, 'Implicit Stereotypes and Prejudice' in GB Moskowitz (ed), *Cognitive Social Psychology: The Princeton Symposium on the Legacy and Future of Social Cognition* (Erlbaum 2001) 359. See also n 4.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

But it was only over the past three decades that scientists dedicated themselves to the systematic study of how implicit and explicit— that is, unconscious and conscious—mental processes interact to determine socially relevant behaviour.[16]

Scientific interest in implicit cognition was reignited in the 1980s by experiments showing that individuals' judgments and behaviour systematically diverge from their consciously endorsed attitudes and beliefs. Early research revealed that explicit attitudes, as measured through self-report questionnaires, did not match individuals' disposition or behaviours toward members of racial out-groups. In such cases, attitudes and interaction could be better explained by unconscious or automatic mental processes, which were neither controlled nor known by the individual.[17]

Social psychologists use the terms *implicit* and *automatic* to designate mental processes that operate outside the realm of conscious awareness and control, distinguishing them from their *explicit* and *controlled* counterparts. The terms *implicit* and *explicit* refer to the perceiver's level of awareness of a psychological process. A process is *explicit* if it can be consciously detected and reported (regardless of whether it was triggered spontaneously), and *implicit* if it cannot be directly inferred through introspective awareness.[18] The terms *automatic* and *controlled* are used to indicate whether a process is intentional, that is, desired by the perceiver, or unintentional.[19]

Psychologists have revealed a wide array of implicit mental phenomena—including implicit memory, self-esteem and perception—and much research has focused on understanding the implicit dimensions of two mental processes that constitute plausible causes of discriminatory behaviour: stereotypes and prejudice. A stereotype is a socially shared set of beliefs about traits that are characteristic of members of a group or category.[20] Common social stereotypes consist, for instance, in characterising women as fragile and weak, or the elderly as dependent or wise. Stereotypes may be positive, negative or neutral—their central aspect is not evaluative but rather descriptive, as they ascribe a

[16] Greenwald and Krieger (n 4) 945–46.

[17] For one of the first demonstrations of implicit intergroup bias, see SL Gaertner and JP McLaughlin, 'Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics' (1983) 46 Soc Psychol Q 23. See also P Devine, 'Stereotypes and Prejudice: Their Automatic and Controlled Components' (1989) 56 J Personality and Soc Psychol 5, on the dissociation between the implicit and explicit components of stereotyping and their impact on behaviour.

[18] AG Greenwald and MR Banaji, 'Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes' (1995) 102 Psychol Rev 4, 4–5; See also TD Wilson, S Lindsey and TY Schooler, 'A Model of Dual Attitudes' (2000) 107 Psychol Rev 101, for an account of the dissociation between implicit and explicit attitudes, and on the persistence of implicit attitudes even after explicit attitudes have been changed.

[19] It is important to state that the dichotomy between implicit and explicit is somewhat artificial, and it would be more accurate to represent mental processes along a continuum which ranges from more implicit to more explicit—see Blair, 'Implicit Stereotypes and Prejudice' (n 15) 361.

[20] MR Banaji and AG Greenwald, 'Implicit Stereotyping and Prejudice' in *The Psychology of Prejudice: The Ontario Symposium*, vol 7 (Taylor and Francis 1994) 55, 58. See also Greenwald and Banaji (n 18) 14.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

trait to a social category and therefore generalise the identity of its members. A prejudice, on the other hand, refers to an essentially evaluative disposition. Prejudices are negative evaluations of a social group and refer to one's unfavourable attitudes toward the group and its members.[21] As a result of prejudices and stereotypes, people will respond more or less favourably to an individual based on their group memberships. When such biases result from implicit attitudes and beliefs, they are referred to as *implicit biases*.[22]

Over the past decades psychologists have discovered that stereotypes and prejudices operate largely on implicit and automatic levels, meaning that individuals rely on stereotypes and prejudiced evaluations without realising or controlling these processes. Even the most low-prejudiced perceivers, those who consciously subscribe to egalitarian principles and condemn discrimination, are not immune to implicit prejudice, and dissociations between explicit and implicit attitudes are common.[23]

B. *Measurement and Pervasiveness of Implicit Bias*

Early research and findings on implicit cognition evidenced the need for a new theoretical framework and innovative empirical tools for the study of stereotyping and prejudice. Traditional self-report measures, that is, measures of explicit attitudes and beliefs, had proven to be poor predictors of intergroup interaction, and new methods had to be developed in order to understand mental processes distorted by intentional dissimulation and introspective inaccessibility.

A series of indirect measurement techniques was developed to this end.[24] Over roughly 30 years of research on implicit mental processes, social psychologists have adapted and deployed well-established scientific methods from the field of cognitive psychology in order to provide 'a solid empirical bedrock for understanding the occurrence of implicit bias'.[25] Many of these methods were used to demonstrate how patterns of knowledge organisation

[21] See Dasgupta (n 15) 145.

[22] Greenwald and Krieger (n 4) 951.

[23] See for example Blair, 'Implicit Stereotypes and Prejudice' (n 15); Dasgupta (n 15); ST Fiske, 'Stereotyping, Prejudice, and Discrimination at the Seam Between the Centuries: Evolution, Culture, Mind, and Brain' (2000) 30 Eu J Soc Psychol 229; and Devine (n 17) suggesting that both high- and low-prejudice subjects produce stereotype-congruent or prejudice-like responses when the subjects' ability to consciously monitor stereotype activation is precluded. According to Devine, '[N]on-prejudiced responses take time, attention, and effort. To the extent that any (or all) of these are limited, the outcome is likely to be stereotype-congruent or prejudice-like responses' ibid 15. Since then, research has demonstrated that even these conditions are not enough to guarantee bias suppression, for time and attention resources may be used to rationalise a bias instead of neutralising or suppressing it. See for example MI Norton and others, 'Casuistry and Social Category Bias' (2004) 87 J Personality and Soc Psychol 817, for a general account of how, given the time and resources, individuals will tend to implicitly mask biased decision-making by deploying more acceptable criteria to justify their choices.

[24] See Jost and others (n 4) for a review of how research on implicit bias emerged out of research on semantic automatic associative links in memory.

[25] ibid 42.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

and information processing in the human mind were related to biases in judgment, decision and behaviour in the absence of intention, awareness or effort.[26] Measures of semantic association were developed to empirically capture implicit social processes (such as stereotypes and prejudice) by measuring the speed and efficiency through which social categories (for example female or male, young or elderly, homosexual or heterosexual) are linked in the mind to evaluative concepts (good or bad, lazy or productive, beautiful or ugly). Experimental variations of this paradigm include serial semantic priming techniques, eye-blink startle responses and cognitive or behavioural inference paradigms.

Amongst these methods, the Implicit Associations Test—IAT—has become the most widely used to study implicit prejudice, stereotyping and intergroup behaviour. The IAT is 'a reaction-time measure that captures the strength with which social groups (and other attitude objects) are implicitly and automatically associated with good/bad evaluations and other characteristics'.[27] The test is used to assess strengths of associations between concepts by observing response latencies in computer-administered categorisation tasks. The IAT is designed in the following way: in an initial block of trials, exemplars of two contrasted categories (eg face images of young or old people) appear on the screen and subjects classify them by pressing one of two keys (eg 'i' for young faces and 'e' for old faces). Next, exemplars of another pair of contrasting concepts are also classified using the same keys (eg key 'i' for words with positive valence and 'e' for words negative valence). Then, in a first combined task, participants categorise the four words using the two keys, each of which has two response options mapped into it (eg 'e' for young or positive and 'i' for old or negative). In a second combined task, the complimentary pairing is used (eg 'e' for young or negative and 'i' for old or positive). The difference in average response latency between these two sets is known as the *IAT effect*, meaning that larger IAT effects reflect stronger implicit associations between concept pairings. In the example above, faster responses for the [young–positive] and [old–negative] task than for the [young–negative] and

[26] According to Jost and others (n 4) 43, '[t]he first demonstrations of implicit stereotyping and prejudice, then, were merely logical extensions of a well-known cognitive principle—namely that knowledge is organised in memory in the form of semantic associations that are derived from personal experiences as well as normative procedures and rules . . . The phenomenon of implicit bias was comfortably assimilated into theories of mundane, workaday principles of human information processing, and it fits the contemporary consensus in the brain and behavioral sciences that an enormous amount of cognition occurs automatically, effortlessly, and outside of conscious awareness.'

[27] Jost and others (n 4) 41; For further descriptions of the IAT and other methods indirect measures, see C Neil Macrae and others, 'Out of Mind but Back in Sight: Stereotypes on the Rebound' (1994) 67 J Personality and Soc Psychol 808, 812; C Neil Macrae, GV Bodenhausen and AB Milne, 'The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping' (1995) 69 J Personality and Soc Psychol 397, 400 (on the Lexical Decision Task—LDT); BA Nosek and MR Banaji, 'The Go/No-Go Association Task' (2001) 19 Soc Cognition 625 (on the Go/No-Go Association Task—GNAT); and AG Greenwald and others, 'Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity' (2009) 97 J Personality and Soc Psychol 17, 18. The Project Implicit Website offers the possibility of taking several varieties of the IAT test, see https://implicit.harvard.edu/implicit/> accessed 24 March 2014.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

[old–positive] task indicate a stronger association of young than of old with positive valence, and the greater the difference between average response latencies the more the participant is biased towards the category 'young'.

An important property of IAT measures is their presumed reliance on associative processes that operate automatically.[28] The implicit nature of the association is reinforced by the speed of the responses, which are assumed to reduce or eliminate the interference of conscious awareness or control, as well as the usage of subliminally activated cues. Different versions of the IAT may be used to measure the strengths of different types of associations—attitudes (concept–valence associations), stereotypes (group–trait associations), self-concepts or identities (self–trait or self–group associations) and self-esteem (self–valence associations).

In a partial literature review, Blair points out that over 30 studies have demonstrated that Whites[29] have relatively strong implicit negative associations with Blacks and other Non-White groups and positive associations with Whites.[30] Evidence for implicit gender bias is also growing, specifically regarding the tendency to implicitly associate men with stereotypically masculine attributes (eg strength, aggression) and women with stereotypically feminine traits (eg dependence, weakness). Furthermore, implicit age bias as well as bias toward a variety of occupational and societal groups has also been widely demonstrated. In many of these studies the mere perception of easily discernible group features (eg skin colour, gender, age-related characteristics) is sufficient to cause the activation of a stereotype, which was then shown to influence judgments of a group member in an unintended fashion, without the perceiver's awareness.[31]

IAT results compiled by Greenwald and Krieger demonstrate the different intensities of implicit bias among different advantaged and disadvantaged groups.[32] While test respondents showed implicit preference for categories such as European American relative to African American, White relative to Asian, young relative to old, and heterosexual relative to homosexual, the intensities of such biases, measured by the *bias index*, were considerably different.[33] The results attest, for instance, that 78.9 per cent of respondents showed implicit preference for young relative to old people, and the bias index for these

[28] Greenwald and others (n 27) 18.

[29] The terminology used in this section to describe social groups may sound objectionable. It is, however, in line with the common usage in the social psychology literature reporting experiments and findings on implicit bias. For the sake of the accuracy of our rendition, we have refrained from changing this terminology.

[30] Blair, 'Implicit Stereotypes and Prejudice' (n 15).

[31] JA Bargh, 'The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects' in S Chaiken and Y Trope (eds), *Dual-Process Theories in Social Psychology* (Guilford Press 1999) 361, 363.

[32] Greenwald and Krieger (n 4) 957, specify that the data reported were obtained from IAT measures in which pleasant and unpleasant words were classified together with items representing advantaged and disadvantaged groups.

[33] The bias index is calculated as the percentage of respondents showing favourability to the advantaged groups minus the percentage showing favourability towards the disadvantaged group. See Greenwald and Krieger (n 4) 955.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

categories is 73 per cent; 69.2 per cent of respondents showed preference for European Americans relatively to African Americans, with a bias index of 57 per cent; and 68.8 per cent showed implicit preferences for heterosexuals over homosexuals, with a bias index of 60 per cent.

Two important results can be inferred from applications of the IAT. Firstly, implicit biases against women, racial minorities, the elderly and other historically underprivileged groups (such as Asians, Muslims and overweight people) are widespread. Second, IAT measures consistently showed that implicit attitude measures reveal far more bias favouring advantaged groups than explicit measures. The former conclusion attests for the pervasiveness of implicit bias, while the latter suggests a dissociation between the implicit and explicit biases.

The findings derived from studies employing the IAT and other semantic association measures have been confirmed by investigations in the neurosciences deploying very different methods to assess the existence and physical manifestations of implicit mental processes. While the IAT seeks access to automatic and implicit processes by limiting the introspective capacity of participants through time constraints, other investigations measure physiological responses—such as skin conductance or levels of brain activity—to assess participants' implicit responses to stimuli.[34] Recently, neuroscientific techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) have enabled researchers to elucidate the neural systems involved in the expression and regulation of implicit attitudes, providing physiological evidence for their existence.[35]

For instance, recent neuroscientific studies have demonstrated that the strength of amygdala activation is predicted by IAT scores but not by conscious measures of race attitudes.[36] Measuring amygdala activity is of special importance because the amygdala 'is critically involved in emotional learning as measured by fear conditioning, a task in which a neural stimulus comes to acquire emotional properties through direct association with an aversive stimulus'.[37] Research along these lines has provided additional evidence for the good construct validity of the IAT by showing that variation in IAT scores are correlated to variations in physiological activities (such as brain activity) that are associated to emotional responses to social stimuli.

C. *From Bias to Behaviour*

Data on the pervasiveness and intensity of implicit intergroup bias would not be so concerning if it were of no consequence to social behaviour, ie, if implicit

[34] D Stanley, E Phelps and M Banaji, 'The Neural Basis of Implicit Attitudes' (2008) 2 Current Directions in Psych Sci 164.

[35] Jost and others (n 4); Stanley, Phelps and Banaji (n 34); E Phelps and others, 'Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation' (2000) 12 J Cognitive Neuroscience 12, 729–38.

[36] Phelps and others (n 35).

[37] ibid 729.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

prejudice and stereotyping did not actually translate into biased responses and discriminatory conduct.[38] The relationship between implicit bias and behaviour is of special importance to anti-discrimination policy. If implicit intergroup bias were proven not to be a plausible cause of discrimination—that is, if people in fact did implicitly stereotype and evaluate based on group membership but for some reason these beliefs and attitudes were irrelevant to their behaviour—then implicit psychological phenomena would be of no significance to law and policy.

Unfortunately, a growing body of evidence suggests that people do make significant social decisions and behave according to their implicit stereotypes and prejudice, without intending to or even being aware of doing so. Furthermore, discriminatory conduct may be rationalised through implicit processes whereby individuals first make a biased decision and then construct objective or bias-neutral criteria to justify themselves.[39] This type of ex-post rationalisation is referred to as casuistry, and poses an additional challenge to the identification and prevention of discrimination.

In a study on the relations among IAT scores, discriminatory behaviour and explicit measures of racial attitudes, McConnel and Leibold revealed that participants who had shown stronger negative attitudes towards Blacks (vs Whites) on the IAT task had more negative social interactions with a Black (vs with a White) experimenter.[40] Larger IAT effect scores predicted greater speaking time, more smiling, more extemporaneous social comments, fewer speech errors and hesitations in interactions with the White (vs Black) experimenters. These verbal and non-verbal responses indicated overall increased friendliness and disposition toward the White interviewer, factors which might be decisive in real-life interactions such as job interviews. According to these researchers,

> There were significant correlations between the IAT and the experimenter's rating of social interaction bias and between the IAT and the judge's molar ratings of social interaction bias. Specifically, as participants' IAT scores reflected relatively more positive attitudes towards Whites than Blacks, social interactions were more positive

[38] C Jolls and CR Sunstein, 'The Law of Implicit Bias' [2006] 94 Cal. L. Rev. 969, 971.

[39] See Norton and others (n 23); EL Uhlmann and GL Cohen, 'Constructed Criteria: Redefining Merit to Justify Discrimination' (2005) 16 American Psychol Science 474 (for experiments demonstrating how criteria of merit are defined in decision-making processes in order to justify biased choices); and NM Lindner, BA Nosek and A Graser, 'Age-Based Hiring Discrimination as a Function of Equity Norms and Self-Perceived Objectivity' (2014) 9(1) PLoS ONE e84752 (for an account of how the presence of equity norms tends to increase casuistry instead of reducing bias in decision-making).

[40] AR McConnel and JM Leibold, 'Relation among the Implicit Association Test, Discriminatory Behavior and Explicit Measures of Racial Attitudes' (2001) 37 J Experimental Soc Psychol 435. The authors found that '[T]he IAT was related to biases in intergroup social interactions. Therefore, researchers can be confident that attitudes assessed by the IAT do relate to intergroup behavior. These findings also suggest that the IAT does assess personal attitudes in that idiosyncratic variability in implicit measures of prejudice was related to behavior. Moreover, the ability of the IAT (unlike explicit measures of prejudice) to predict several specific biased social behaviors as assessed by independent observers is consistent with the claim that implicit measures of attitudes are especially predictive of behavioral leakage', 440.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

> toward the White experimenter that toward the Black experimenter as assessed by both trained judges and by the experimenters themselves.[41]

Other studies provide evidence of even more concerning implications of implicit racial stereotyping. Correll, Park, Judd, and Wittenbrink, for instance, designed a 'shoot/don't shoot' framework to investigate the influence of race on decisions to shoot or not shoot potentially armed targets, and discovered participants had a pronounced bias to shoot Blacks.[42] In the experiments, participants perform a videogame task in which they encounter armed and unarmed targets that are either Black or White. Participants are instructed to shoot armed targets and to indicate 'don't shoot' in response to unarmed ones. Participants proved faster and more likely to shoot Black targets, and were also faster and more likely to indicate 'don't shoot' for Whites.[43] The authors explain these results by arguing that the concept of 'danger' is a component of the stereotypical characteristics attributed to Blacks, and as the experiment's design triggered this stereotype, it led participants to react to Black targets as if they were dangerous.

Due to increased interest in how implicit associations impact on behaviour, many other studies using an IAT attitude measure have also included a measures of social behaviour that are theoretically expected to be correlated to attitude or stereotype measures. These studies include measures of warmth, friendliness and discomfort during interracial and intergroup interaction. Analysis of the data then determines whether individual differences in implicit attitudes or stereotypes measured in the IAT correlate with differences in behaviour.[44]

In 2009, Greenwald, Poehlman, Uhlmann, and Banaji conducted a meta-analysis of 122 such studies in order to estimate the average predictive validity effect size of IAT and self-report measures.[45] In order to analyse the data, effect sizes were separated into nine domains based on similarities among criterion measures: White–Black race interactions, other intergroup interactions, gender/sexual orientation, consumer preferences, political preferences, personality traits, alcohol and drug use, clinical phenomena and close relationships.

Based on their compilation of empirical data, the authors of this meta-study drew a conclusion that is especially relevant for the present research. Namely, the meta-study revealed that for the samples with criterion measures involving Black–White interracial interaction and other intergroup behaviour—such as behaviour toward groups defined by ethnicity, age, or weight—the predictive

[41] ibid 439.
[42] J Correll and others, 'The Influence of Stereotype on Decisions to Shoot' (2007) 37 Eur J Soc Psychol 1102.
[43] ibid 1102–3.
[44] Greenwald and Krieger (n 4) 954.
[45] AG Greenwald and others, 'Understanding and Using the Implicit Association Test III: Meta-Analysis of Predictive Validity' (2009) 97 J Personality and Soc Psychol 17.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

validity of IAT measures significantly exceeded that of self-report measures.[46] In other words, when the experiment involves inter-group interaction, the attitudes and beliefs explicitly reported by the individual are not good indicators of how that person is likely to act. In these cases, the way a person behaves is more closely correlated to her implicit attitudes than to explicit ones.

These data suggest that discriminatory behaviour is driven by implicit biases. This presents a fundamental challenge to how anti-discrimination law addresses this type of behaviour. These results provide systematic evidence that the predictive validity of implicit attitude measurements outperform measures of explicit attitudes in socially sensitive domains such as stereotyping and prejudice.[47]

While scientists have widely documented the influence of implicit bias on judgment and behaviour, social category information has become taboo both in discourse and decision-making. People are therefore under constant internal and external pressure to appear unbiased and objective both to themselves and to others.[48] Taking this tension into account, researchers have studied how individuals deal with their own biases and found that decision-makers tend to engage in casuistry and find ex-post rationalisations for their biased choices.[49] This is done by reconstructing or adapting the criteria used in decision-making in a way that will justify a biased decision in a neutral manner. Even more importantly, these ex-post rationalisation processes are also largely implicit and therefore difficult to detect and control.

In an experiment reproducing a hiring scenario, Uhlmann and Cohen demonstrate that after making employment decisions based on gender stereotypes, participants were likely to construct ostensibly objective criteria to justify their biased choices.[50] The process occurs implicitly—that is, without awareness of the participant—and such ex-post rationalisation had the effect of concealing biases under a cape of neutrality and rationality. The actor himself does not realise he is discriminating, and in the presence of neutral criteria to back up the decision, discrimination becomes harder to spot and prove.

It has also been shown that certain social cues that increase pressure on the decision-maker to be objective will increase casuistry instead of reducing bias. In another hiring-scenario experiment, Lindner, Nosek, and Graser demonstrated that in the presence of an equity norm—that is, an equal opportunity

[46] For interracial behaviour, aggregate ICC ($r_{ICC} = .24$) was significantly greater than aggregate ECC ($r_{ECC} = .12$), and for the 'other intergroup' category, which included behaviour toward groups defined by ethnicity, age, or weight, aggregate ICC ($r_{ICC} = .20$) was also greater than aggregate ECC ($r_{ECC} = .12$). Greenwald and others (n 45) 24.

[47] Stanley, Phelps and Banaji (n 34).

[48] Norton and others (n 23) 817–18.

[49] See Norton and others (n 23); and Uhlmann and Cohen (n 39).

[50] Uhlmann and Cohen (n 39).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

statement prohibiting discrimination—participants retrospectively reported their hiring decisions as relying more on bias-neutral criteria (in the specific experiment this referred to the applicant's expertise) and less on social category information (in the specific case this concerned the age of the applicant).[51] At the same time, the participants' decisions remained just as biased in the presence of the equal opportunity statement as before. Hence, the statement just had the effect of increasing casuistry.[52]

In sum, it has become clear that implicit bias is not 'just' a cognitive phenomenon. It is also of far-reaching social significance. Moreover, there is increasing evidence that individuals, even if alerted to their potential bias, may not be capable of correcting its behavioural effects.

4. *Implications for Law and Policy*

We have seen that over the last three decades research in implicit social cognition has yielded a wealth of insights. Even if the stream of research is still in flux and many open questions remain,[53] the state of the art already provides solid foundations for reconsidering the policies that address bias and discrimination. It is high time for policy makers to draw on these insights and put their current policy mix to the test of behavioural realism.[54]

A. *The Problems with Traditional Anti-Discrimination Law*

We stated in the introduction that anti-discrimination law has largely failed to solve the problem at which it is targeted. The problem, we may recall, is that social inequality is being continuously reproduced by individual decisions that systematically disadvantage certain groups. Traditional anti-discrimination law

[51] Lindner and others (n 39).

[52] ibid 6.

[53] The questions to be derived from research into implicit social cognition refer to the basic traits of different forms of bias as well as the effectiveness and durability of specific strategies for bias reduction. First, there is a need to better understand the specificities of different types of biases. Research has demonstrated that ageism, sexism, and racism are biases with different features—some are more intense than others, more entrenched or malleable, etc. The compilation presented by Hamilton and Krieger suggests that biases in favour of European Americans (v African Americans) or young people (v the elderly) are significantly more pervasive than biases against people who are homosexual or overweight. See Greenwald and Krieger (n 4) 957. Furthermore, research has also suggested that different biases are responsive to different types of control strategies. See, for example N Dasgupta and AG Greenwald, 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals' (2001) 81 J Personality and Soc Psychol 800, for evidence suggesting that exposure to counter-stereotypical imagery had a stronger impact on implicit racism than on implicit ageism, and that the effects of bias control interventions tend to diminish over time. Scientists and policy makers still need to know more about the determinants of such durability, as well as how reliable these strategies are when different types of bias are considered. For policy makers, it is also fundamental to know how different decision-making procedures influence the occurrence of bias and its impact on behaviour. Instituting collective decision-making procedures that force actors to justify their decisions might be an effective strategy for bias reduction.

[54] For a nuanced concept of such realism and a discussion of its prospects in policy debates, see A Benforado and J Hanson, 'Seeing Bias: Discrediting and Dismissing Accurate Attributions' in Jon Hanson (ed), *Ideology, Psychology, and Law* (OUP 2011) 453–98. With regard to its prospect in discourses on law (and economics), see A Benforado and J Hanson, 'Backlash, The Reaction to Mind Sciences in Legal Academia' in Hanson (ed) ibid.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

addresses this problem by trying to restrain individual decision-takers from taking into account certain characteristic features of the disadvantaged group.

This approach has long been known to face compliance problems. Individuals who deliberately base their decisions on the very group features which they are meant to ignore might be inclined to conceal this motivation. Much attention has thus been devoted to the issue of how to prove such discriminatory motivation or its absence, and on whom the burden of proof should rest.

These concerns are no doubt plausible, and so are the responses. But they address only part of the compliance problem. The insights from social psychology suggest shifting our attention from deliberate discrimination to implicit bias as another and maybe even more important reason for the inefficacy of anti-discrimination law. The reported lines of research show that implicit bias is marked and pervasive, that it affects behaviour and, most importantly, that it is hard to overcome even for individuals who would be willing to comply with a command to that extent.

In some cases, and arguably the easy ones, such inability to avoid one's own biases is owed to the restrictive circumstances of a specific situation. As the shooting example[55] illustrated, lack of time might force an individual to rely entirely on her automated cognitive operations. Similarly, a person might be too distracted or exhausted to mobilise the resources necessary to identify and counteract an implicit bias.

It is important to acknowledge the cognitive strain that self-monitoring entails. For this may well pose a relevant obstacle to anti-discrimination law compliance in practice. Nonetheless, this problem does not seem insurmountable. Most of the important decisions that are targeted by anti-discrimination law could probably be shaped in a way that allows for sufficient deliberation.

In many cases, however, the effects of implicit bias are not confined to the moment of decision-taking, but might affect many prior cognitive operations. To take another example from the employment context, consider a promotion decision. At the time when the superior evaluates the individual records and takes the decision, she might have interacted with the candidates for a long time, collected impressions, assigned tasks, written assessments etc. At all of these instances, her perceptions and actions might have been biased, and it would seem practically impossible retrospectively to correct the accumulated bias.

Admittedly, the promotion context is specific in that it encompasses a particularly long period of—potentially biased—interaction. But it is evident that there are many other decisions that involve a chain of perceptions and interactions[56] and should hence be conceptualised as a process that is liable in

[55] See Correll and others (n 42).
[56] See McConnel and Leibold (n 40).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

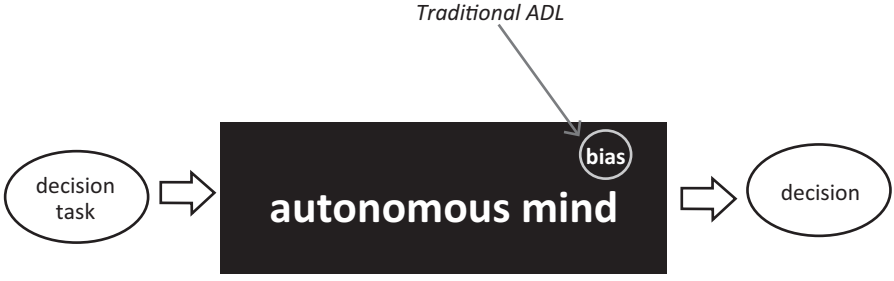

Figure 2. The Human Mind, According to Traditional Anti-Discrimination Law

principle to the same kind of bias accumulation as illustrated by the promotion example.

More intriguingly still, it seems that the self-correction of bias might fail even in the absence of such practical constraints. Recall the research on equity norms[57] which suggests that the effects of implicit bias are not even mitigated by alerting individuals beforehand. Or recall the evidence on casuistry.[58] This suggests that although the mind does seek to maintain congruence between its explicit views and its decisions, it does this by retrospectively adjusting the reasons for the biased decision rather than by correcting the bias itself. Moreover, this ex-post adjustment process, just like the initial bias, occurs without introspective awareness or even control.

Taken together, the evidence shows that there are various factors that can make it hard, if not impossible, to abide by a legislative command to disregard one's own biases. It thus seems that traditional anti-discrimination law is based on too simplistic a concept of the cognitive processes involved. From its perspective, the human mind still features as a black box (see Figure 2), a mysterious entity that produces decisions when faced with the respective tasks, and that should be respected as autonomous unless under exceptional circumstances outside interference is warranted—as in the case of discriminatory decisions when the biased considerations, and only these, are eliminated from its operations by way of narrowly targeted legal intervention.

Although still far from unravelling the mystery of the human mind, the findings reported above allow for a more refined conceptualisation of the relevant cognitive processes. This helps explain the shortcomings of the interventions of anti-discrimination law. And it can also guide the search for more effective policy intervention.

We may at this point recall that our focus in this article is on decisions by individuals and in individual cases, not on rules or policies that apply to

[57] See Lindner and others (n 39).
[58] See Norton and others (n 23); and Uhlmann and Cohen (n 39).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

multiple cases and may be adopted in collective procedures. Bias, to be sure, may become operative even in decisions of this latter kind. But the ways it does this are possibly more complex and certainly not as fully understood as in the case of individual decisions. This is one reason why it is worth emphasising the focus of our present discussion. Another reason is that anti-discrimination laws pertaining to individual decisions differ significantly from the respective rules that apply to other decisions. In a nutshell, the latter have in a long-term perspective come to be more concerned with discriminatory effect than with discriminatory intent.[59] Such a turn towards an objectivised measure could certainly help overcome enforcement problems due to implicit bias.

B. *Alternative Policy Options*

We have seen that the insights from social psychology cast additional doubt on the viability of anti-discrimination laws. The lesson, in short, was that bias occurs without the affected person's awareness, and that in many cases, it cannot be overcome even if it is brought to that person's attention. Under these circumstances, anti-discrimination law can hardly be effective.

Hence, alternative policies should be sought and integrated into a more comprehensive mix of anti-discrimination policies. Having identified the neglect of implicit bias as a central problem of anti-discrimination law, we will in this section develop a tentative systematisation of alternative policy interventions that can be pursued alongside or instead of traditional anti-discrimination law.[60] We will explore the strengths and limitations of each option in light of the empirical research presented in previous sections. Our goal is not to provide a comprehensive assessment or ranking of the policy options and their relative viability or efficacy, but rather to map out the range of anti-discrimination policies that could be used to address the problem of implicit bias and point to the contextual conditions under which we might expect them to be more or less adequate. Figure 3 may serve as an initial illustration of these alternative options to intervene.

The diagram shows a simple schematic depiction of the process of individual decision-taking. It displays a decision task that is to be processed by a human mind, which in turn produces a decision. At the centre, the human mind is still a 'black box', albeit with a few more specifications. It incorporates a time dimension so that processes that require little time are on the left side, while

[59] The core of this evolution is the development of the concept of indirect discrimination; for a thorough treatment from a comparative perspective, see Fredman (n 1) 177.

[60] Even if traditional anti-discrimination law may have other shortcomings regarding compliance and enforcement, it is our exclusive aim in this article to focus on the inability of this policy strategy to tackle implicit bias.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

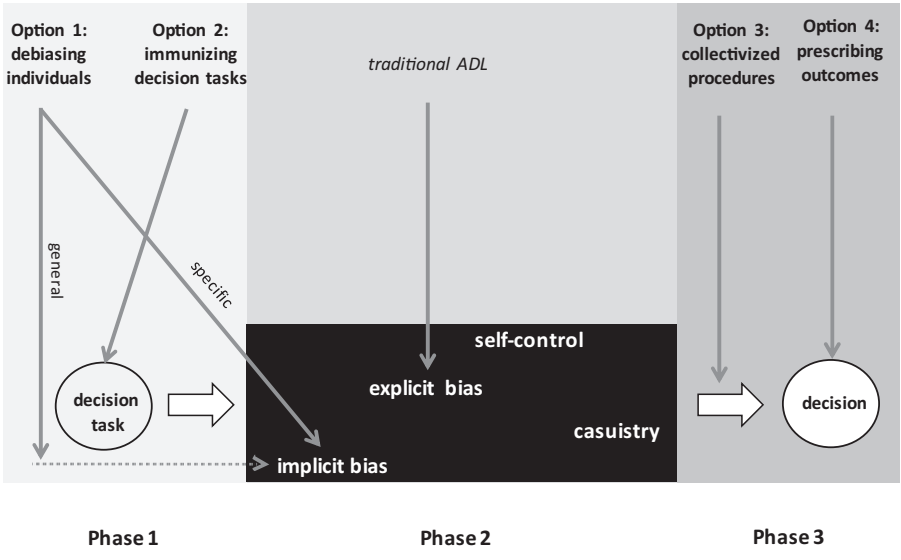

Figure 3. Traditional Anti-Discrimination Laws (ADL) and Alternative Policy Options

those that require more time are located more towards the right. Also, the vertical axis roughly reflects the level of awareness. Thus, implicit processes are at the bottom, those that require some degree of awareness are positioned higher up.

The diagram distinguishes three phases in which intervention can take place: before, during, and after the individual decision is formed. As has been pointed out above (see section 4A), many decisions will involve more than the one-stage operation that our simplistic schema depicts. Typically, decision-takers will face a task, deal with it and revisit it various times before they form and ultimately communicate a decision. Hence, the basic sequential structure that our schema suggests may be misleading. Reality is, of course, more complex. This will have to be borne in mind when considering the possibilities of intervention.

The policy options are symbolised by the thin arrows. Apart from traditional anti-discrimination law, we may distinguish four broad types of intervention that we will discuss below. It may be noted that while traditional anti-discrimination law is concentrated on Phase 2, the alternative interventions mostly fall outside of that phase and take place before or after the actual decision-taking. This is in recognition of the above evidence that a targeted correction of the mental process is more complicated and liable to failure than has hitherto been assumed.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

(i) Option 1: Debiasing through the law

The set of interventions that we have labelled Option 1 does include measures that affect Phase 2. They are part of what some authors have recently referred to as 'debiasing through law'.[61] Their starting point is the insight that implicit biases are not static or isolated, but in constant interaction with multiple variables including the perceiver's state of mind, his motivations, and contextual stimuli. To the extent that these variables can be manipulated, implicit bias may become malleable.[62]

Based on the recent and mostly incipient research on the malleability of stereotypes and prejudices, most debiasing through law strategies seek to 'mitigate objectively measurable bias by producing environmental conditions that alter the strength of associations between social category and attitude and attribute'.[63] In trying to reduce implicit bias, these strategies may either be directed at a specific decision situation (thus falling in Phase 2), or address more generally the overall level of bias in a given community.

Policies of the latter kind have long been employed. They are typically motivated by what is referred to as the 'social contact hypothesis', that is, the expectation that the overall prevalence of bias may be reduced if people get to know and interact more often with members of a group that is subject to bias. Accordingly, any measure that seeks to diversify the individual experience of such social contact may be viewed as a general debiasing strategy and hoped to indirectly reduce the effect of implicit bias in individual decision-taking. Along these lines, Jolls and Sunstein argue that many elements of current equality-oriented policies may be interpreted—and justified—as direct debiasing mechanisms.[64] Take, for example, traditional anti-discrimination law. To the extent that it is effective in preventing explicitly biased decision-making, it may increase population diversity, which in turn reduces levels of implicit bias.[65] Laws prohibiting hostile environments are also likely to have debiasing effects, and by promoting more diverse environments, affirmative action programmes can also be thought of as direct debiasing strategies.

Far less common is the idea of shaping a specific decision situation so as to counteract implicit bias on part of the decision-taker(s). But experiments on bias malleability offer some guidance in this regard. Research has shown, for

[61] For an account of bounded rationality and the use of debiasing strategies in situations where heuristics and bounded rationality are likely to play an important role, see C Jolls and CR Sunstein, 'Debiasing Through the Law' (2006) 35 JLS 199; for an account of implicit bias and a review of why and how debiasing strategies can be used to reduce it, see Jolls and Sunstein, 'The Law of Implicit Bias' (n 38).

[62] For partial reviews of the literature on malleability if implicit bias, see IV Blair, 'The Malleability of Automatic Stereotypes and Prejudice' (2002) 6 Personality and Soc Psychol Rev 242; and MJ Monteith and CI Voils, 'Exerting Control Over Prejudiced Responses' in GB Moskowitz (ed), *Cognitive Social Psychology: The Princeton Symposium on the Legacy and Future of Social Cognition* (Erlbaum 2001) 375.

[63] Kang and Banaji, 'Fair Measures' (n 2) 1110.

[64] Jolls and Sunstein, 'The Law of Implicit Bias' (n 38) 981.

[65] See ibid 981; and see also N Dasgupta and S Asgari, 'Seeing is Believing: Exposure to Counterstereotypic Women Leaders and its Effects on the Malleability of Automatic Gender Stereotyping' (2004) 40 J Experimental Soc Psychol 642.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

instance, that some sorts of motivation—for instance the motivation to maintain one's self-image[66] or to preserve valuable social relationships[67]—can lead individuals to suppress negative stereotypes and prejudice, and that contextual factors such as exposure to counter-stereotypical imagery[68] and thoughts[69] may, under certain circumstances, reduce the level of implicit bias. In light of these findings, the mere presence of images of counter-stereotypical personalities during a decision-taking process—such as a photograph of a successful and popular member of a minority group in the room where a job interview is conducted—may help reduce bias (on both parts, by the way, because a minority candidate may be affected by negative self-stereotyping which might also be reduced this way). Similarly, one may consider more far-reaching interventions along these lines, such as requiring the presence of

[66] See L Sinclair and Z Kunda, 'Reactions to a Black Professional: Motivated Inhibition and Activation of Conflicting Stereotypes' (1999) 77 J Personality and Soc Psychol 885. Sinclair and Kunda demonstrated that when people are motivated to preserve their self-image in face of an external threat, they are likely to protect themselves through the activation of negative stereotypes regarding the threatening other. See also SJ Spencer and others, 'Automatic Activation of Stereotypes: The Role of Self-Image Threat' (1998) 24 Personality & Soc Psychol Bulletin 1139, demonstrating that people whose self-image has been threatened may automatically activate negative stereotypes to make themselves look better or discredit a disliked evaluator. See generally Blair, 'Implicit Stereotypes and Prejudice' (n 15) for a review of different variables that influence the activation and strength of bias.

[67] See for example JA Richeson and N Ambadi, 'When Roles Reverse: Stigma, Status, and Self-Evaluation' (2001) 31 J Applied Soc Psychol 1350, for an account of how the relative status of the perceiver can moderate the extent to which she inhibits automatic prejudice. In their study, White participants were informed they would be working with an African American student. Some were told they would evaluate their partner's performance during interaction (superior role), others were instructed to cooperate and get along with their partners (equal-status role) and others were informed they would be evaluated on their performance (inferior role). Participants assigned to a superior role in the interracial interaction produced higher levels of automatic prejudice than participants assigned to an equal-status role, and participants who were assigned to a subordinate role exhibited the least amount of prejudice.

[68] See N Dasgupta and AG Greenwald, 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals' (2001) 81 J Personality and Soc Psychol 800 for evidence on the impact of exposure to counter-stereotypic group members of automatic prejudice. In the first experiment, participants were shown images of either admired Black Americans and disliked White Americans or admired White Americans and disliked Black Americans, or to non-racial stimuli. Participants then completed a race IAT both immediately after the exposure and 24 hours later. The researchers found that individuals exposed to positive Black group members produced less automatic bias against Blacks (IAT effect = 78 ms, $d = 0.58$), compared to participants who had been exposed to negative group members (IAT effect = 176 ms, $d = 1.29$) or to non-racial stimuli (IAT effect = 174 ms, $d = 1.15$). Furthermore, the moderation continued to be significant 24 hours after the exposure. A second experiment replicated the effect for the moderation of automatic prejudice, this time using age categories. Again, exposure to pro-elderly exemplars yielded a smaller automatic age bias effect (IAT effect = 182 ms, $d = 1.23$) than exposure to pro-young exemplars (IAT effect = 336 ms, $d = 1.75$). There was no control group and the IAT test was not applied again 24 hours after the counter-stereotypic exposure. Here again counter-stereotypic stimuli seemed to significantly reduce intergroup bias, but not eliminate it. Also worth noting are the large IAT effects obtained in the age-bias test when compared to the race-bias evaluation.

[69] See IV Blair, JE Ma and AP Lenton, 'Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery' (2001) 81 J Personality and Soc Psychol 828, demonstrating that thoughts, through a form of self-priming, moderated implicit stereotypes. A group of participants was instructed to imagine a strong woman, her attributes and abilities; another groups was instructed to imagine a vacation scene in the Caribbean. Those who imagined the strong woman registered a significantly lower level of stereotyping in the IAT—for the neutral imagery group, the reaction time difference between the schema-consistent and schema-inconsistent blocks was 95 milliseconds; for the counter-stereotypic imagery group, the difference was 24 milliseconds. Notice that the interventions did not eliminate the bias, but reduced it significantly.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

'debiasing agents', that is, counter-stereotypical minority members, throughout the decision-taking process.[70]

Promising though these potential interventions certainly are, we know at this point far too little to effectively control and direct implicit cognition. And even if future research were to close these gaps, it is already apparent that there are limits to this approach. There has not been any experiment so far in which contextual and subjective manipulations such as those described above were able to completely eliminate bias. Also, there is indication that the effects of bias manipulation fade quickly and that not all kinds of bias are equally responsive. Tackling these problems will be amongst the biggest challenges faced by proponents of debiasing strategies in the future.[71]

It seems that intervening in the 'black box' remains a challenge, and that successful strategies of specific debiasing are very context-sensitive. Designing such an intervention requires detailed knowledge of the setting of the targeted decision, and it also requires a considerable degree of control over this setting so as to avoid interferences by any unforeseen context factors. It may well be feasible to meet these conditions in specific cases. But it would seem very demanding to design debiasing policies that are apt for general use and could be made mandatory. Furthermore, since not all bias is eliminated by the researched debiasing strategies, it would generally seem reasonable to combine them with other measures as part of a broader anti-discrimination policy mix (see section 4C(i)). In sum, specific debiasing strategies seem to be a promising addition to the anti-discrimination policy basket, may already be deployed in individual situations, and certainly deserve more attention and research in future.

(ii) Option 2: Immunisation strategies

There is another set of policy measures, labelled Option 2 above, which intervenes before the actual decision-taking process is initiated. They seek to immunise the decision situation against the effects of bias by concealing bias-inducing information from the decision-taker and thereby *impeding* the formation of implicit bias. There are various ways of doing this.

One way is to shield off the relevant features of an applicant from the perception of the decision-taker. Instead of assuming that implicit biases can be reduced and controlled, such 'shielding' strategies part from the assumption that individuals will inevitably be biased when exposed to certain stimuli, and that procedural rules should implement strong safeguards to ensure that outcomes will not be 'contaminated' by biases. Looking once again at the employment context, and at hiring decisions more specifically, blind auditions in symphony orchestras have come to be the paradigmatic case in point. Since

[70] Kang and Banaji, 'A Behavioral Realist Revision of "Affirmative Action"' (n 2) 1109.

[71] BA Nosek and R Riskind, 'Policy Implications of Social Cognition' (2012) 6 Social Issues and Policy Rev 113, 129.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

the 1970s, a series of reforms has been instituted in the selection processes of symphony orchestras in the USA in order to ensure more egalitarian opportunities for candidates.[72] One of these reforms consisted of using physical screens during the audition to conceal the candidate's identity and guarantee impartiality in the selection process. Goldin and Rouse found that:

> [T]he screen increases—by 50 percent—the probability that a woman will be advanced from certain preliminary rounds and increases by several fold the likelihood that a woman will be selected in the final round. By the use of the roster data, the switch to blind auditions can explain 30 percent of the increase in the proportion female among new hires and possibly 25 percent of the increase in the percentage female in the orchestras from 1970 to 1996.[73]

This is promising evidence that such immunisation strategies have significant impacts on bias avoidance.

Further applications of that very principle include banning certain information (such as gender, age etc) from the candidates' CV, or prescribing extended reliance on phone interviews so the candidates' visual appearance cannot play a role. It may be noted that in both of these examples, practical considerations may require that the sensitive features of the candidates be revealed at some point still within the decision-taking process. But the impact of potential bias will at least be reduced if it cannot affect the earlier stages of that process.

In the same vein, a slightly weaker immunisation strategy may consist in educating decision-takers about the effects of implicit bias so that they frame their own working procedures in a way that reduces its impact. Also, it may make sense especially in large firms to delegate the respective decision tasks to individuals who display a low level of implicit bias. To identify the appropriate individuals, the IAT or similar tests might serve as a selection tool.

Finally, a decision can also be immunised against bias by prescribing reliance on objective criteria such as, in the employment context, degrees, grades, entry exams, etc.[74] It should be noted, though, that there will often remain significant discretionary leeway in the assessment of individual criteria and the relative weight assigned to them. Complete formalisation of a decision with regard also to these dimensions is very demanding, and it may hence be viewed as undesirable. This applies in particular when not all relevant aspects of the

[72] C Goldin and C Rouse, 'Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians' (2000) 90 Amer Econ Rev 715, 716.

[73] ibid 738.

[74] In one experiment on gender stereotyping in hiring scenarios, Uhlmann and Cohen found that having decision-makers commit previously to the use of determinate hiring criteria reduces the level of bias against women for a typically male job position. According to the authors, '[m]en who had not committed to hiring criteria prior to disclosure of the applicant's gender gave more favorable evaluations to a male applicant for police chief than to a female applicant. By contrast, men who had committed to criteria prior to disclosure of the applicant's gender gave equivalent evaluations to the male and female applicants. Our research thus demonstrates the efficacy of a method to reduce job discrimination: the establishment of standards of merit prior to the review of candidates'. See Uhlmann and Cohen (n 39) 478–79.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

decision situation can be foreseen, such as, in an employment setting, the size and quality of the pool of applicants.

(iii) Option 3: Procedural strategies

Yet another set of measures, labelled Option 3 above, relies on the implementation of decision-making procedures which *may detect and cancel out* the effects of implicit bias. The basic idea is to require the decision and its reasons to be shared with other individuals before it becomes final. Before entering such communication, the individual will typically have at least begun to process the decision task herself, so that implicit bias may have become operative already. This is why Figure 3 depicts Option 3 in Phase 3. Nonetheless, as the decisions relevant in this context will typically be multistage, the required communication (and possibly interaction) with others is likely to affect the individual decision-taking process and not just the handling of its outcome.

There is, again, a vast variety of such procedural measures. They range from the mere requirement that a written account of the decision be produced for supervision, via the obligatory consultation of certain persons (such as minority representatives), to the delegation of the decision in all its stages to a committee (possibly pluralistic in its composition). Such procedures may serve multiple purposes, such as to enhance the legitimacy of the decision or its transparency, to broaden its informational base, or simply to distribute the related work. Counteraction of individual bias may well be among these purposes, although it is by no means clear that interaction would per se reduce implicit bias. It is conceivable, to be sure, that any such communication, written or oral, helps detect bias by requiring that the individual decision be rationalised. But this outcome is not guaranteed. In fact, the communication may well work the opposite way, eg in settings when an influential committee member manages to spread her implicitly biased view among other committee members.

Much will therefore depend on the specific procedures in question. Designing them appropriately, however, is difficult. This is all the more true as very little is known so far about how bias plays out in such communication generally, and in group interactions in particular. With a view to the critical position that collectivised decision-taking procedures occupy in a democracy, closing this gap would be an important desirable for future research.

(iv) Option 4: Prescribing outcomes

The fourth and last of set of interventions identified in Figure 3 seeks to bypass the impacts of implicit bias by making prescriptions for the outcome of the targeted decisions. Again, one can distinguish different measures that come under the heading of Option 4. The outcome might, in theory, be dictated for each individual decision. It is much more common, though, for such

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

prescriptions to relate to a number of decision outcomes and to require just a fraction of them to comply with the intended result. This is what quotas do.

Such regulation may be qualified in various ways. It might, for example, only apply if applicants meet a certain threshold of qualification. Or it might allow for a delay in meeting the prescribed targets. Also, enforcement structures can differ significantly. Because of their generalised design, quotas cannot typically be enforced via individual claims but require some collective monitoring and sanctioning system instead.

Relatedly, the same outcome that quotas seek to achieve for members of disadvantaged groups may also be sought by mandating preferential treatment for them within the decision-taking procedure. Such intervention may target the very end of the decision-taking process, prescribing that in cases of equal qualification a member of a disadvantaged group be preferred. Or it may, especially in cases of highly formalised procedures, afford an assessment bonus to members of a disadvantaged group. This latter intervention, however, might influence but will not determine the outcome (and hence would fall between the categories used here).

(v) Interim summary

Table 1 may serve as a summary of the foregoing discussion. It provides an overview of the options available in anti-discrimination policy. It also lists the main advantages and disadvantages of these policies.

C. *Towards a New Mix of Anti-Discrimination Policies*

Our overview of policy measures other than traditional anti-discrimination law has illustrated the multitude of alternative policy options that may be deployed to address implicit bias. At the same time, it has also shown that the available policy measures tackle different dimensions of bias at different phases of decision-making. Given their specificities and limitations, some of these policies may be effective only under certain conditions or with regard to specific types of biases (that is, debiasing strategies). Others may be viewed as undesirable for other reasons, as will be explored below.[75]

Also, we have seen how the doubts about the efficacy of traditional anti-discrimination law have grown. But this does not imply that this kind of policy had to be discarded altogether. For one, we did not go as far as to claim that it could not work at all. Indeed, traditional anti-discrimination law seems indispensable for the prevention and prosecution of explicit forms of bias—which unfortunately are still present in society. Furthermore, such laws represent a normative statement which is not only valuable in itself, but which

[75] See section 4C(ii).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

Table 1. Summary of Anti-Discrimination Policy Options

	Traditional anti-discrimination law	Debiasing strategies	Immunisation strategies	Procedural strategies	Outcome prescribing strategies
What it seeks to do	Prevent conscious bias in decision-making tasks.	Reduce levels of implicit bias.	Inhibit the formation of implicit and explicit bias.	Cancel out the effects of implicit bias.	Bypass explicit and implicit bias.
How it does it (operational mechanism)	Prohibiting certain criteria from being taken into account in decision-making.	Producing environmental conditions that alter the strength of associations between social category and attribute or attitude.	Concealing bias-inducing information from decision-taker.	Requiring the inter-subjective communication of the decision and its reasons before it becomes final.	Proscribing outcomes of targeted decisions.
Examples of concrete interventions	Anti-discrimination norms mandating that protected criteria (age, gender, colour, weight, etc) ought not to play a role in decision-making.	Introduction of debiasing agents in decision-making processes, exposure to counter-stereotypical imagery.	Shielding off protected features of an applicant from the perception of the decision-taker; banning protected criteria (age, gender, colour etc) from candidates' CVs.	Requiring the submission of a written account of the decisions to supervisors, delegation of decisions to committees.	Affirmative action policies, most notably mandatory quotas for disadvantaged groups.
Advantages	Is narrowly tailored so as to leave room for autonomous decision-making; might have general debiasing effects.	Tackles the very existence of implicit bias.	Is effective in preventing the onset of implicit and explicit bias.	Helps in the detection and correction of bias without being too intrusive on decision-making.	Ensures that the outcome of decision-making will not be determined by explicit or implicit bias.
Disadvantages	Does not tackle—at least not primarily—the implicit dimensions of bias; may increase casuistry.	Does not completely eliminate bias; effects of manipulation may fade quickly; not all types of bias are equally responsive; difficult to control.	Practical considerations normally require that protected traits be revealed during the decision-making process.	The impacts of procedural strategies on implicit bias have not been sufficiently researched and desired outcomes may not be reached.	Does not, at least primarily, address the existence of bias; criticised for being overly intrusive upon decision-making (but see section 4C(ii)).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

may ultimately also bear important debiasing effects upon individuals and hence be classified as what we termed a *general* debiasing strategy earlier.[76]

Taken together, and quite unsurprisingly, this means that we cannot distil a clear blueprint for reform from any of the above. We will still have to select from a wide array of policy options, in fact from an even wider one than before. We will still be uncertain about the promise of any of these measures. And we will still have to adjust any strategy that we may choose to the variable conditions of the time and locality in question. The battle against the continuous reproduction of social inequality due to individual decisions—or more succinctly, the pursuit of third layer equality-oriented policies—has not become any less intricate than it has been so far.

But this is not to downplay the impact that implicit social cognition may have on practical policy-making. Not only is this a relatively recent and rapidly growing field of research—many of the unsettling questions and speculations that it has produced so far may soon turn into solid answers and insights—but much more importantly for now, there are a number of practical lessons that can be inferred at this stage, and that indeed suggest the adoption of a new policy mix. Moreover, we may also find that the evaluative matrix by which we assess and choose the relevant policy interventions may be shifted in light of the insights from implicit social cognition.

(i) Lessons for a more comprehensive basket of anti-discrimination policies

A first and very straightforward lesson that we draw from the research on implicit social cognition is that reliance on what has been the central tool in this policy field, traditional anti-discrimination law,[77] is likely to decrease. In light of the evidence presented above and given the search for innovative policy tools designed to tackle implicit bias, it is likely that traditional anti-discrimination law will become but one element amongst many others, and that other policy options—debiasing strategies, immunisation measures, outcome prescription—will gain relative importance within an increasingly complex and comprehensive anti-discrimination policy basket.

A second and related lesson is that the instruments available to tackle different forms of bias—from explicit to implicit and everything in between—possess different strengths and limitations, and hence should be taken as complementary strategies in a comprehensive anti-discrimination policy basket. Preventing explicit bias, reducing levels of implicit bias, preventing the formation of the latter if possible and, if not, cancelling out or bypassing its effects are all potentially important forms of tackling the complex phenomenon of discrimination. These mechanisms—associated, respectively, with traditional anti-discrimination law, debiasing strategies, immunisation strategies,

[76] See section 4B(i).

[77] For a presentation of traditional anti-discrimination law and its historic contextualisation, see sections 2B and 2C.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

procedural strategies and affirmative action—tackle different aspects of the social problem that third-layer equality-oriented policies seek to overcome. What we have tried to show in this article is that a policy focused only on addressing the explicit dimension of bias misses a large part of the story.

Furthermore, evidence from implicit social cognition suggests—and this is our third lesson—that policy makers must take seriously the difficulties and pitfalls of intervening in the 'black box', and that they must be aware of the obstacles that stand in the way of controlling the mental processes of individual decision-takers. In a context in which implicit bias is pervasive and we are still uncertain about how to impede its formation and to cancel out its effects, the case to be made in favour of outcome-oriented policies may be stronger than it has been considered to date. We shall elaborate on this last point in the next section.

(ii) Towards new justifications: putting autonomy in perspective

In addition to introducing a new array of policy options to the field of anti-discrimination policy, the research programme on implicit social cognition is also likely to change the way we evaluate the range of available policy instruments and justify our respective choices.

As pointed out before,[78] the main objection against third layer equality-oriented policies[79] appears to be that by interfering with individual decision-taking, they infringe upon personal autonomy, which in turn is considered a constitutive element of individual liberty and as such enjoys a heightened degree of legal protection. However, not all third layer equality-oriented policies interfere with individual decision-taking to the same extent. So those which entail a relatively modest infringement would, from this perspective, be preferable over other, more intrusive devices.

Hence the predominance, as we suggested before, of traditional anti-discrimination laws among third layer equality-oriented policies. For this device would appear relatively un-intrusive, especially if compared to affirmative action policies. Clearly, it should be more respectful of individual autonomy to simply prohibit a certain consideration in the decision-taking process than to determine (part of) its outcome.

There is nothing wrong with this reasoning as such. It is, however, rooted in behavioural assumptions that are typically left implicit and hence underdeveloped in policy debates. The respect of autonomy presumes that individuals are indeed capable of autonomous perception, judgment and action, that is, that they can perceive 'objectively' any decision-task before them, distinguish and

[78] See sections 2A(iv) and 2B.

[79] For a more detailed account of principled objections against antidiscrimination policies, see Alfinito Vieira and Graser, 'The Case Against the Case Against Affirmative Action' (n 12).

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

choose freely which goals to pursue, and act rationally towards the realisation of these goals.

But as we saw in the previous section, the cognitive underpinnings of social interaction in general and of discriminatory behaviour in particular are difficult to reconcile with the notion of personal autonomy as it underlies public policies in this field. We discriminate without even being aware of it—not always, to be sure, but most of us most of the time. What is more, we have, even if made aware of such discrimination, only very limited capacity to control and correct our behaviour accordingly. And all of this applies regardless of—and often even despite—our deliberate intentions and consciously endorsed values.

The question we are left with in light of this evidence, then, is what and whose autonomy are we talking about. The presumed addressee of such respectful policies, the 'autonomous mind', occupying the black box in Figure 2, seems to operate in a very peculiar way: it is not aware of its own biases, would, if it were, object to them—and yet, if alerted to them, it would retrospectively rather adjust its own criteria than correct its biased decisions.

Now, this entity is no doubt still capable of autonomy. But the question is: is it worthy of respect? It would seem that in the field of anti-discrimination policies, we need to distinguish different mechanisms that are operative in the human mind:[80] there are the deliberate, conscious parts—that we typically associate with the rational and possibly also moral self—and there are the unconscious parts, which we have seen to be especially determinative of our social interactions in general, and of discriminatory behaviour in particular. Paying respect to autonomy in anti-discrimination law typically means to have the unconscious part rule. Intervention, by contrast, infringes upon that part's autonomy. But it is an autonomy that bears little weight as it is grounded neither in rational deliberation nor moral reflection on part of the individual. And indeed, intervention may in these cases well support what our conscious self would endorse.

Coming back to our starting point, this argument erodes the ground for preferring traditional anti-discrimination law over more 'intrusive' policies, such as 'Option 4' policies, most notably affirmative action and quotas. Where implicit bias is likely to operate, there is little basis to respect individual autonomy. Hence the insights from implicit social cognition are likely to weaken the case for traditional anti-discrimination law in many contexts, and to strengthen the one for affirmative action policies.

[80] For a recent and influential exposition of such non-unitarian thinking about 'the' mind see D Kahneman, *Thinking, Fast and Slow* (Farrar Straus & Giroux 2011), who uses the metaphors of 'System I' and 'System II' to contrast different cognitive modes.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

It should be noted that our concluding argument does not imply a challenge to the notion of personal autonomy altogether, nor to the related concepts of a free will and a rational mind. All of them, to be sure, may require a major overhaul when viewed in light of the increasing knowledge on human cognition.[81] But this is far beyond the scope of this article, and the present discussion is but one small component of this larger debate.

Downloaded from http://ojls.oxfordjournals.org/ at MPI Study of Societies on April 1, 2016

[81] See most notably the wealth of pertinent reflections collected at www.lawandmind.com by the Project on Law and Mind Sciences at Harvard Law School. Such an overhaul would have to take account of the multiple cognitive fallacies that the human mind is liable to (for an overview on the state of research see Kahneman (n 82)), and it might have to move from 'fallacies'—which presuppose an agreed standard of rationality—to a notion of relative situational aptitude; for this approach, see G Gigerenzer, *Rationality for Mortals* (OUP 2008) 18–19, who proposes a context-dependent concept of 'ecological rationality' that assesses the rationality of cognitive heuristics by the degree to which they are adapted to a certain environment.