Book Section:
Holroyd, J. and Picinali, F. (2022) Implicit bias, self-defence, and the reasonable person. In: Lernestedt, C. and Matravers, M., (eds.) The Criminal Law's Person. Hart Publishing / Bloomsbury Publishing, pp. 167-192. ISBN: 9781509923748.

https://doi.org/10.5040/9781509923779

Implicit Bias, Self-Defence, and the Reasonable Person

Jules Holroyd & Federico Picinali

ABSTRACT: The reasonable person standard is used in adjudicating claims of self-defence. In US law, an individual may use defensive force if her beliefs that a threat is imminent and that force is required are beliefs that a reasonable person would have. In English law, it is sufficient that beliefs in imminence and necessity are genuinely held; but the reasonableness of so believing is given an evidential role in establishing the genuineness of the beliefs. There is, of course, much contention over how to spell out when, and in virtue of what, such beliefs are reasonable.

In this chapter, we identify some distinctive issues that arise when we consider that implicit racial bias might be implicated in the beliefs in imminence and necessity. Considering two prominent interpretations of the reasonable person standard, we argue that neither is acceptable. On one interpretation, we risk unfairness to the defendant - who may non-culpably harbour bias. On another, the standard embeds racist stereotypes. Whilst there are formulations of the defence that may serve to mitigate these problems, we argue that they cannot be avoided in the presence of racist social structures.

Introduction

In this paper, we explore the impact that the findings from empirical psychology about cognitive bias should have on the concept of 'person' in the criminal law.[1] More precisely, our focus is on the reasonable person standard. As we will show, the finding that we are beset by cognitive biases might be thought to undermine, on at least some occasions, the extent to which we are reasonable. What implications does this have for the application of the standard?

Our enquiry is narrowed in two ways. First, whilst the reasonable person standard appears at various points in criminal law, we focus our attention on its role in the claim of self-defence in both US and English law. Second, we focus on implicit racial biases; in particular, on the related phenomena of weapons bias, shooter bias, and perceptions of aggression. This is because these phenomena are directly relevant to the beliefs and actions at issue in claims of self-defence. It is an open question that we leave here unaddressed whether other cognitive biases pose similar challenges for this or other parts of the law.

We proceed as follows. First, we outline the phenomena of weapons bias and shooter bias. Second, we articulate and modify a much-discussed test-case of self-defence. This allows us to import the phenomena of racial bias into a hypothetical scenario in which the claim of self-defence is at issue. Thirdly, we consider how the defence, as formulated in US and English law respectively, would apply to our test-case – with particular attention to the different role that the reasonable person standard plays in the two jurisdictions. In the course of our discussion of the defence, we employ a distinction introduced by Dan-Cohen between rules of conduct (directed to the public and regarding

[1] For another recent exploration of the impact of findings about implicit bias on conceptions of responsibility in criminal law, see N Lacey, 'Socializing the Subject of Criminal Law: Criminal Responsibility and the Purposes of Criminalization' (2016) 99 *Marquette Law Review* 541, 551-3.

how to act) and decision rules (directed to officials and regarding how to respond to a putative violation of a legal norm). This distinction helps us to better articulate the tensions that arise when the reasonable person standard is applied to individuals harbouring implicit biases. We argue that plausible ways of construing the reasonable person standard in light of implicit racial bias face either the charge of being unfair towards the biased individual, or of stigmatising the group targeted by the bias. The bottom line is that insofar as the reasonable person serves as a normative ideal, this ideal is not one that can be met easily whilst our agency is embedded in unjust social relations.

1. Unreasonable Persons and Biased Beliefs

A huge research programme in social psychology has revealed that individuals frequently display a range of cognitive biases. These include confirmation bias – the disposition to more readily believe evidence consistent with one's prior beliefs; familiarity bias – the disposition to make preferential judgements of things with which we are familiar; anchoring bias – the disposition to be swayed in our judgements by an arbitrary fixed point, to which our subsequent judgements are anchored. All of these biases are grounded in modes of automatic thinking that are often useful: they provide fast automatic cognitive short-cuts that enable us to avoid cognitively demanding processing. For example, confirmation bias means that we don't have to deliberate from scratch about the full set of evidence available, for or against that belief, on each occasion we have to form a new belief. But whilst often useful, these 'habits of cognition' may sometimes distort our reasoning.[2] A feature of these cognitions which makes us particularly ill-placed to identify these distortions is that their operation is often difficult to detect (perhaps because they operate for the most part automatically); it is fast (because automatic), and so is difficult to exert control over. These cognitive biases are often called 'implicit cognitions'.

Certain kinds of implicit cognition have garnered a great deal of attention from psychologists and philosophers, and with good reason.[3] These are the cognitions that encode information about social identity, such as race, gender, age, sexuality and associated characteristics. Our social cognitions may encode problematic associations that link stereotypical characteristics with social group membership, despite our explicit disavowal of those stereotypes. For example, our cognitions might associate men more strongly with leadership qualities than women.[4] Or we might associate white people more

[2] See Patricia Devine, Patrick Forscher, Anthony Austin, William Cox, 'Long-term Reduction in Implicit Race Bias: A Prejudice Habit-breaking Intervention' (2012) 48 *Journal of Experimental Social Psychology*, 1267-78.

[3] See John Jost, Laurie Rudman, Irene Blair, Dana Carney, Nilanjana Dasgupta, Jack Glaser and Curtis Hardin, 'The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of 10 Studies that no Manager Should Ignore' [2009] *Research in Organisational Behaviour* 29, 39-69; Calvin Lai, Maddalena Marini, Steven Lehr, Carlo Cerruti, Jiyun-Elizabeth Shin, Jennifer Joy-Gaba and Arnold Ho, 'Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions' (2014) 143 *Journal of Experimental Psychology: General* 1765.

[4] Virginia Valian, *Why So Slow? The Advancement of Women* (MIT Press 1999).

strongly with intellectual constructs than black people.[5] These associations have been widely detected by a number of indirect measures, such as the Implicit Association Test (IATs).[6] To discover that there are aspects of our cognition that we may repudiate is itself troubling; but worse, they appear to have a role in producing discriminatory judgements and behaviours.[7] For example, these biases have been found to correlate with differential evaluations of the same CV, and differential judgements of suitability for hiring, when the only variable is the gender or race of the person to whom the CV belongs (as indicated by the name at the top of the CV).[8] Also, medical practitioners harbouring implicit racial biases make different prescription recommendations for patients reporting the same symptoms but who differ with respect to race.[9] And white individuals' 'microbehaviours' – non-verbal indicators of tension or discomfort – have been found to differ in inter-racial interactions, where greater signs of tension and lesser attentiveness are displayed by white interlocutors. [10] Insofar as one maintains that gender or race should be irrelevant to the evaluation of the quality of an applicant's materials; or irrelevant to the disposition to prescribe treatment; or irrelevant to the quality of one's interactions with others, these biases can be said to have a distorting effect on judgement and behaviour.

The research literature on implicit bias is vast (see Jost for a comprehensive, but already out of date literature survey[11]), and dismayingly indicates that we are at risk of these distortions far more often than we might otherwise have supposed. Here we want to focus on two particular instances in which implicit biases might impact on beliefs or actions. We then turn to the relevance of these cases for the reasonable person standard in criminal law.

1.1 Biased Perceptual Judgements

A number of studies have revealed 'weapons bias': the tendency to associate weapons

---

[5] David Amodio and Patricia Devine, 'Stereotyping and Evaluation in Implicit Race Bias: Evidence for Independent Constructs and Unique Effects on Behaviour' (2006) 91 *Journal of Personality and Social Psychology* 652.

[6] Indirect measures access cognitions via means other than self-report, which is notoriously unreliable. See Greenwald Nosek and M Banaji, 'The Implicit Association Test at Age Seven: A Methodological and Conceptual Review' in J Bargh (ed), *Automatic Processes in Social Thinking and Behaviour* (Psychology Press 2007) 265-92, for a review of indirect measures, including the Implicit Association Test (IAT).

[7] Of course, there are other important explanatory considerations, such as social structure and institutional design (see Sally Haslanger, 'Distinguished Lecture: Social Structure, Narrative and Explanation' (2015) 45 *Canadian Journal of Philosophy* 1, 1-15).

[8] John Dovidio, Samuel Gaertner, Kerry Kawakami and Gordon Hodson, 'Why Can't We Just Get Along? Interpersonal Biases and Interracial Distrust' (2002) 8 *Cultural Diversity and Ethnic Minority Psychology* 88.

[9] Alexander Green, Dana Carney, Daniel Palin, Long Ngo, Kristal Raymond, Lisa Iezzoni and M Banaji, 'Implicit Bias Among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients' [2007] 22 *Journal of General Internal Medicine* 1231, 1238.

[10] Dovidio (n 8).

[11] Jost (n 3).

more readily with black males.[12] When asked to identify ambiguous objects, experimental participants are more likely to identify an object as a weapon when they have been primed with a black male's face, than in the condition in which the primes are white male's faces. The hypothesis is that individuals more strongly associate with weapons – particularly guns – black males than white males. Whilst not statistically supported, these associations may be entrenched by media presentations that promote racial stereotypes. The problematic outcome is that this has an impact on individuals' perceptual judgements of material objects. This finding garners further support in 'shooter bias' tasks.[13] Experimental participants are presented with scenes in which black or white males are depicted holding ambiguous objects (which, in fact, are guns, mobile phones, drink cans) and told that their time-limited task is to press 'shoot' or 'don't shoot' depending on whether the individual before them is armed. The finding (that has been replicated not only in US, but also UK populations) is that individuals more readily make the error of shooting an individual who is not armed when that individual is black. The hypothesis, again, is that individuals form mistaken perceptual judgements of ambiguous objects.[14] Perceptual judgements under normal circumstances are thought to provide reason for belief; but in these cases, perception provides misleading evidence. What misleads, here, is our own cognition – implicit racial biases. This is a phenomenon that philosophers have called 'cognitive penetration', whereby one's own prior cognitions taint one's perceptual evidence.[15]

To clarify the problem, let's take the belief *g*: 'that there is a gun'. In the case in which an individual is unarmed, and implicit racial bias distorts perceptual judgement, the belief *g* would be formed more readily if the individual is black. Many of us rightly balk at the idea that we might form such perceptual judgements on the basis of race, and take such beliefs to violate an important moral ideal (which might be cashed out in terms of respect, or equality). But the belief would also be in bad epistemic shape, since it would violate a fairly uncontroversial epistemic norm, such as that 'perceptual judgements influenced by distortive biases do not provide justification for belief'.[16] The difficulty, of

---

[12] Keith Payne, 'Weapon Bias Split-second Decision and Unintended Stereotyping' (2006) 15 *Current Directions in*
Psychological Science 287.

[13] Jack Glaser and Eric Knowles, 'Implicit Motivation to Control Prejudice' (2008) 44 *Journal of Experimental Social Psychology* 164.

[14] A competing hypothesis is that individuals' motor responses are readier for action when faced with a potentially armed black male than a potentially armed white male. Keith Payne, Yujiro Shimizu and Larry Jacoby, 'Mental Control and Visual Illusions: Toward Explaining Race-biased Weapon Misidentifications' (2005) 41 *Journal of Experimental Social Psychology* 36, found some support for this, but on other occasions, they found that the error was one of mistaken perceptual judgement.

[15] Susanna Siegel, 'Cognitive Penetrability and Perceptual Justification' (2012) 46 *Noûs* 201. Siegel discusses the problems that the 'cognitive penetration' of belief pose for the epistemic status of perceptual evidence, including cognitive penetration of perception by implicit bias.

[16] In a much lower stakes example of this: perceptual judgements of length in Muller-Lyer illusions do not provide justification for belief. However, consider that epistemic norms may be violated also if the belief turns out to be true. If founded in bias, the belief would not be appropriately evidentially sensitive (cf. Alex

course, is in knowing when one's perceptions are distorted in this way.[17] The point at this stage is not to make a judgement about the reasonableness or otherwise of making these mistakes (more on this later), but rather to point out that the resulting beliefs can readily be identified as defective: as violations of an uncontroversial epistemic norm.[18] Yet many of us may well harbour such weapons biases (indeed, fair-minded undergraduate students at the University of Sheffield manifested these biases).[19]

Consider also the following set of studies that focused on distorting biases, but this time on perceptions of aggression. The first focuses on aggression as expressed in black and white faces. Hugenberg and Bodenhausen (2003) found that white individuals more readily identified faces as expressing anger and hostility when the faces were black, rather than white.[20] Individuals who showed greater anti-black implicit bias (i.e. stronger associations – measured with IATs – between black people and negative terms, than white people and negative terms) were more ready to judge an ambiguous black face as hostile. Again, such perceptual judgements appear to be distorted by implicit racial biases, and so any resulting belief (about the degree of hostility manifested) lacks adequate perceptual justification.[21]

Madva, 'Why Implicit Attitudes are (probably) Not Beliefs' (2016) 193 *Synthese* 2659; Neil Levy, 'Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements' (2015) 49 *Noûs* 800.

[17] See Jennifer Saul, 'Implicit Bias, Stereotype Threat and Women in Philosophy' in Katrina Hutchinson and Fiona Jenkings (eds), *Women in Philosophy: What Needs to Change?* (Oxford, Oxford University Press 2013) 39-60, for the worry that insofar as we are ill-positioned to detect when we are influenced by biases, this generates a radical kind of skepticism that she calls 'bias-induced doubt'.

[18] See J Holroyd and K Puddifoot, 'Implicit Bias and Prejudice' in Miranda Fricker, Peter Graham, David Henderson, Nikolaj Pedersen and Jeremy Wyatt (eds), *The Routledge Handbook of Social Epistemology* (Taylor and Francis 2019) (forthcoming) for various specifications of this epistemic norm, under all of which the norm is violated by beliefs based on implicit biases. However, some have argued that base rate information provides useful support for our beliefs and, especially, that implicit biases may provide evidential support for belief by encoding such information (see Gendler 2011). This supposes that implicit biases are the kind of mental state that can represent precise statistical data. But it is unlikely that implicit biases accurately represent such information, insofar as they are crudely associative (see K Puddifoot, 'Dissolving the Ethical/Epistemic Dilemma over Implicit Bias' *Philosophical Explorations (*forthcoming)) and are not appropriately evidentially sensitive (see Madva (n 16), Levy (n 16)). It is for these reasons that we endorse the claim that the phenomenon of statistical discrimination is distinct from implicitly-biased behaviour (see also RM Blank, M Dabady and CF Citro (eds), *Measuring Racial Discrimination* (Washington DC, The National Academies Press 2004).

[19] R Scaife, J Holroyd, T Stafford and A Bunge, 'The Effects of Moral Interactions on Implicit Racial Bias' (MS).

[20] Hugenberg and Bodenhausen 2003

[21] This recent study corroborates an older one: in Birt Duncan, 'Differential Social Perceptions and the Attribution of Intergroup Violence: Testing the Lower Limits of Stereotyping of Blacks' (1976) 34 *Journal of Personality and Social Psychology* 590, participants were asked to evaluate ambiguous behaviour (described as 'a shove'). When participants observed this behaviour perpetrated by a black against white individuals, it was judged to be an act of violent hostility. When perpetrated by white against black individuals, it was evaluated as 'playing around'. Note that this study was conducted in 1976, when racial attitudes may have been somewhat different from those prevailing now, so it is less clear that these conclusions generalise to us, here and now.

We have reason to believe that such implicit biases and their distortive effects are widespread. It is likely that many of us have biases of this kind. Simply possessing such biases is widely thought to be non-culpable: they are frequently formed on the basis of exposure to associations in our environment, independently of whether individuals endorse or subscribe to the problematic stereotypes that they encode.[22] And, some have argued that since it may be difficult to be aware of such implicit biases, until such awareness is gained we are not culpable for failing to take steps to rid ourselves of these biases.[23]

It is tempting to immediately conclude that individuals who harbour implicit biases and form perceptual judgements and beliefs under the influence of implicit bias (hereafter 'bias-based beliefs'), are necessarily unreasonable. Certainly there is nothing to be said to condone such patterns of cognition, and they are clearly defective and damaging. In some pre-theoretical sense, it seems quite clear that such biases generate unreasonable patterns of inference and resultant belief. However, the task ahead of us is to consider how the reasonable person standard, as it is invoked in criminal law, could or should deal with such patterns of inference and belief.

2. Racism and Self-Defence

In this section, we draw attention to a case – 'the case of the mistaken racist'– which has received much attention in the scholarly literature.[24] For our purposes, this case provides a helpful model that we can modify to envisage the role that implicitly-biased action might play in cases of self-defence. The primary role of this case is to animate the implicit biases introduced in the previous section. Garvey (2008) presents the following example, modelled on the Goetz case[25] (we paraphrase):

The case of the mistaken racist: G is riding public transport, when he is approached by

---

[22] Patricia Devine, E Ashby Plant, David Amodio, Eddie Harmon-Jones and Stephanie Vance, 'The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice' (2002) 82 *Journal of Personality and Social Psychology* 835; Saul (n 17).

[23] Natalia Washington and Daniel Kelly, 'Who is Responsible for this?' in Brownstein and Saul (eds), *Implicit Bias and Philosophy* (Oxford, Oxford University Press 2016). Full disclosure: one of us has argued that we may be blameworthy for discriminatory behaviour due to implicit bias (J Holroyd, 'Responsibility for Implicit Bias' (2012) 43 *Journal of Social Psychology* 274; J Holroyd, 'Implicit Bias, Awareness and Imperfect Cognitions' (2014) 33 *Consciousness and Cognition* 511; J Holroyd and D Kelly, 'Implicit Bias, Character and Control' in Jonathan Webber and Alberto Masala (eds), *From Personality to Virtue* (Oxford, Oxford University Press 2016); J Holroyd, 'What do we Want from a Model of Implicit Cognition?' (2016) 116 *Proceedings of the Aristotelian Society* 153). However, if interpersonal blame is not a suitable analogue for state punishment (J Holroyd, 'The Retributive Emotions: Passions and Pains of Punishment' (2010) 39 *Philosophical Papers* 343), and the conditions for responsibility in interpersonal relations differ from those necessary for criminal liability, the arguments made there cannot be carried over to this context.

[24] See Stephen Garvey, 'Self-defence and the Mistaken Racist' (2008) 11 *New Criminal Law Review: International and Interdisciplinary Journal* 119; George Fletcher, *A Crime of Self-Defense: Bernhard Goetz and the Law on Trial* (University of Chicago Press 1988).

[25] *People v Goetz* 68 NY 2d 96 (1986).

> two young black males. From their perceived demeanor and repeated demand for money, in conjunction with his past experience of being mugged on the subway – also by black men – G forms the belief that he faces a threat of serious physical harm, and perhaps mortal danger. He believes he may be killed and that lethal force is needed to defend himself. G pulls out a concealed weapon, shoots and kills one of the men, before fleeing (later to turn himself in).[26]

Much of the discussion surrounding the case (including Garvey's description of Goetz and G as racist) supposes that G endorses repugnant racist stereotypes, or even feelings of racial animosity, and that these are behind the beliefs that he is in mortal danger and that lethal force is required. We hold that beliefs based in explicit racism are obviously unreasonable.[27] However, importing into this scenario the empirical findings from the previous section, we face a rather different set of concerns when considering the question of reasonableness:

> *The case of the mistaken implicitly-biased individual:* B is riding public transport, when he is approached by two young black males. Like many of us, B has implicit racial biases, which inform his perceptual judgements both of the degree of hostility manifested, and of the presence of a weapon in the hand of one of the men. On the basis of the distorted perceptual judgements of the young men's demeanor and behaviour, B forms the belief that he faces a threat of serious physical harm, and perhaps mortal danger. He believes he may be killed and that lethal force is needed to defend himself. B deploys potentially lethal force to repel the perceived danger.

In jurisdictions such as the US or England, we would expect individuals such as B to be charged with attempted-murder or murder, and to try to establish that they acted in self-defence.[28] If they succeed, the outcome of the trial is an acquittal. A successful claim of self-

---

[26] See Garvey (n 24) 123-5. This case differs in some important respects from the real Goetz case, in which all four of the men were wounded, one of whom after what seemed like pre-meditated action. Significantly, Goetz had also previously been attacked and wounded on the subway, which was appealed to in support of the claim that his belief in imminent attack was reasonable.

[27] Though see Garvey (n 24) for an extended discussion of whether this renders the defence unavailable. However, consider that the presence of explicit racist beliefs raises doubt as to whether the agent acted in order to defend herself or for some other unlawful purpose. *Cf* Andrew Ashworth and Jeremy Horder, *Principles of Criminal Law* (7th edn, Oxford University Press 2013) 122.

[28] NB: under English law the defendant only has a to bear an evidential burden with respect to the issue of self-defence. It will then be for the prosecution to disprove self-defence beyond a reasonable doubt. We are leaving aside the question as to whether a partial defence (e.g., of provocation or loss of control) may be available and may be put forward in such cases. However, to the extent that such a defence relies on an assessment of reasonableness the considerations that we make here may apply to it as well. To be sure, the 'new' English defence of loss of control does not include a reasonable person standard, but appeals to the notion of a 'normal degree of tolerance and self-restraint' (see the Coroners and Justice Act 2009, s 54(1)(c)). A test based on this notion seems problematic if implicit biases are the norm. Note also that whilst we focus on murder or attempted murder, in English and US law a claim of self-defence can be made for any intentional

defence may show that the individual conduct was in fact *justified*: that the use of such force was necessary to avert more serious criminal harm to themselves (or to others). Alternatively, the defence may serve to *excuse*. For the time being, though, our focus is on the justificatory role of self-defence, as it is in this context that the problems raised by implicit bias are most evident. In section 6 we return to the excusatory role of self-defence.[29] The availability of the defence in the US and in England rests on whether the defendant (B) can show that she has met a distinctive set of conditions. Each set of conditions makes reference to standards of reasonableness to which individuals are held, though these standards are somewhat differently deployed in each of the two jurisdictions. So, it is instructive for our purposes to consider how the qualifying conditions for self-defence would deal with implicit bias in each case.

Notably, the few treatments of the research findings on implicit bias in relation to self-defence have operated on the assumption that the bias-based belief is unreasonable. For example, Cynthia Lee takes seriously the observation that 'most individuals would be more likely to "see" a weapon in the hands of an unarmed Black person than an unarmed White person'.[30] But she is mainly concerned that if this observation is true of jurors, they 'may also be more likely to find that an individual who says he shot an unarmed Black person in self-defence because he believed the victim was about to kill or seriously injure him acted reasonably'.[31] She finds this problematic as she assumes that this belief is unreasonable. Attention to the judgement of juries is of course important. But Lee's assumption requires careful consideration. Whilst we have no interest in defending the claims that such biased-based belief is reasonable, we must consider further precisely what characterisation of the reasonable person standard supports the conclusion that such a belief is indeed unreasonable, and whether that characterisation is independently defensible.

2.1 A Useful Device for Thinking about Self-Defence

Deploying a heuristic device introduced by Meir Dan-Cohen,[32] we can say that self-defence does double duty both as a rule of conduct and a decision rule. Rules of conduct are addressed to the general public, conveying information about what behaviours are permissible or prohibited. In the case of self-defence, the rule conveys the message that the use of force is permissible, subject to certain conditions being met. In contrast, decision rules are directed to officials in the criminal justice system (judges, jurors etc.), and convey instructions regarding how to deal with individuals charged with violating a rule of conduct. In the case of self-defence, the instructions are that individuals should not be

---

harm to the person.

[29] Another variable that we consider there is the distinction between full and partial defence.

[30] Cynthia Lee, 'Making Race Salient: Trayvon Martin and Implicit Bias in a Not Yet Post-racial Society' (2013) 91 *North Carolina Law Review* 1555, 1584.

[31] ibid, 1585.

[32] Meir Dan-Cohen, 'Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law' (1984) 97 *Harvard Law Review* 625.

punished if their use of lethal force meets certain conditions. One and the same legal norm may be intended and/or perceived as both a rule of conduct and a decision rule – as in the case of self-defence. However – as Dan-Cohen argues – this 'double duty' can create tensions and mixed messages: decision rules may generate 'side-effect' messages pertaining to conduct, for example.[33] The distinction between the two types of rules – or the two dimensions of the same rule – is helpful to us: it brings to light some of the tensions that emerge in the ways that implicit bias may be dealt with by different versions of the reasonable person standard, as deployed in the claim of self-defence in US and English law.

3. US Law and the 'Reasonable-Belief Rule'

In US law, for the defence of self-defence to be available to individuals such as B, it is not sufficient that B believed that there was imminent lethal danger and that potentially lethal force was required to avert the threat.[34] The relevant beliefs must also be *reasonable*, irrespective of whether they are in fact true. Following Baron, the question of reasonableness may be framed in terms of whether a reasonable person might believe as B did, and so act as B did, in the same circumstances.[35]

As has been much discussed, the answer to this formulation of the question depends in part on what features are built into the reasonable person – what are the 'relevant particulars' of the individual that we suppose are shared with the reasonable person.[36] In other words, what background beliefs, dispositions, cognitive processes should we hold fixed in deciding whether the individual in that circumstance believed reasonably? The key question in our case concerns whether the reasonable person might be supposed to share with B – and indeed, with many of us – the implicit biases that distorted perceptual judgement, causing the beliefs that threat was imminent and force required. Let us consider some (more or less) promising ways of articulating the

[33] Dan-Cohen discusses the case of duress, where a norm intended to guide the decision-maker is likely to be understood by citizens as a norm guiding behaviour – thus undermining the force of important criminal law prescriptions. See Dan-Cohen (n 32) 632-4.

[34] We are well aware that speaking of 'US law' as if it were a single and coherent legal system is at best imprecise. Not only may the federal criminal law and the state criminal law differ; the penal codes of the different states present important differences as well. Theoretical works sometimes refer to the Model Penal Code as providing an approximate indication of the state of the criminal law in the US. Whether this strategy is appropriate or not, the Model Penal Code is not our focus here, given that it does not require that the beliefs relevant to self-defence be reasonable. As far as federal law is concerned, our point of reference is the case *United States v Peterson* 483 F 2d 1222 (1973). At the state level, instead, we refer to the Penal Code of the state of New York – which was the relevant code in the Goetz case. Both Peterson and the New York code require reasonableness as a feature of the beliefs that are relevant to self-defence.

[35] See Marcia Baron, 'The Standard of the Reasonable Person in Criminal Law' in RA Duff, Lindsay Farmer, SE Marshall, Massimo Renzo and Victor Tadros (eds), *The Structures of Criminal Law* (Oxford, Oxford University Press 2011) 15-6. Baron argues that the better framing of the standard is this, rather than in terms of what *the* reasonable person *would* have believed or done, since the phrasing in the text makes clearer that there is latitude: that reasonable people may disagree or come to different conclusions.

[36] ibid 17.

reasonable person standard,[37] what they might indicate with respect to this question, and what the further implications of such construals might be.[38]

3.1 The Reasonable Person as the Ordinary or Typical Person

One way of construing the reasonable person is as the ordinary or typical person.[39] According to this interpretation, if the average or ordinary person would believe as B did, then B's belief is reasonable. As we have seen, since implicit biases are pervasive, most likely the ordinary or typical person would have the cognitive biases to which B is susceptible, and so would have the distorted perceptual judgements that lead to the beliefs that B had. On this construal of the reasonable person, B believed reasonably, and so the defence of self-defence would be available; thus, B's use of potentially lethal force would be justified.[40]

This is an uncomfortable line of reasoning. Recall that we noted that the distorted perceptual judgement did not provide adequate justification for the belief that threat was imminent and lethal force required. Yet, this process may be entirely typical and ordinary. If the reasonable person standard is construed as the ordinary person, and what they would believe, then what it is reasonable to believe may – as in this case – lack justification. For those who see the reasonable person as the – in the relevant respect – justified person this is untenable.[41] And, as many authors have noted,[42] it is manifestly true that what is ordinary or typical is often unjustified, defective, or even repugnant in some ways. Baron references the 1896 US Supreme Court ruling that racially segregated travel

---

[37] The construals of the reasonable person that we address here are prominent in the literature, but by no means the only interpretations possible. As will become apparent, though, the tension that we discuss in the paper does not hinge on the particular interpretations of the reasonable person standard at issue, but rather on the fact that in the context of implicitly-biased behaviour both the choice of affording the defence and that of denying it are unpalatable.

[38] To the reader: if your interest is only in English law, these discussions of unreasonableness may seem irrelevant. But as we will see, assessments of reasonableness are not wholly absent from the English law concerning self-defence – so the following discussion will pay off when we come to consider English law in section 5. See also the previous note.

[39] Mark Kelman, 'Reasonable Evidence of Reasonableness' (1991) 17 *Critical Inquiry* 798, 800; Baron (n 35) 26-30; John Gardner, 'The Many Faces of the Reasonable Person' (2015) 131 *Law Quarterly Review* 563, 564-6; Mayo Moran, *Rethinking the Reasonable Person: An Egalitarian Reconstruction of the Objective Standard* (Oxford, Oxford University Press 2003) 13-6.

[40] Cf. Jodi Armour, 'Race Ipsa Loquitur: Of Reasonable Racists, Intelligent Bayesians and Involuntary Negrophobes' (1994) 46 *Stanford Law Review* 781, 787 ff, strongly criticising the concept of the 'reasonable racist'. According to the 'reasonable racist standard', reasonableness exclusively depends on typicality. Assuming that the 'typical American' believes that 'blacks are prone to violence' – hopefully, an assumption that was already dubious at the time when the article was written – it may well be judged reasonable for defendants such as G or B to harbour the relevant beliefs in imminence and necessity. Notice that the argument of the reasonable racist is framed in terms of beliefs rather than implicit biases. In this respect, it is weaker than the argument offered here – given that the following considerations of fairness towards the defendant (see section 4.1) do not arise in the case of racist beliefs.

[41] Gardner (n 39).

[42] Kelman (n 39); Baron (n 35); Gardner (n 39); Armour (n 40); Moran (n 39).

arrangements were deemed reasonable, a judgement made with reference to ordinary and established conduct.[43] So much speaks against taking what is reasonable to be coextensive with what is ordinary or typical. This holds in the case of implicit bias as much as in the case of more familiar beliefs or preferences.

3.2 The Reasonable Person as Having 'Reasonable Basis' for Belief

Baron draws attention to the construal of 'reasonableness' that the New York Court of Appeals gave in justifying their decision in the Goetz case.[44] The crucial part is the Court's claim that a determination of reasonableness must be based on the defendant's circumstances, which encompass 'any prior experiences he had which could provide a reasonable basis for a belief that another person's intentions were to injure … him or that the use of deadly force was necessary.'[45] On this view, the reasonable person is modelled as having reasonable bases for her beliefs. Therefore, the defendant's past experiences are built into the reasonable person insofar as those experiences provide a reasonable basis for belief. Accordingly, the reasonable person may have false or mistaken beliefs, informed by inferences from the defendant's past experiences, so long as those experiences provide a reasonable basis. In the Goetz case, his past experiences of assault by black males were considered and it was asked whether they should be part of the relevant circumstances in the sense articulated above. The conclusion was that they should not, since they were not found to provide a reasonable basis for his belief that *these* individuals pose an imminent threat – this sort of race-based inference was not considered a reasonable basis for belief.

In the case of implicit bias, then, we might say that whilst the ordinary person may well believe as B did, B wouldn't be reasonable insofar as her belief is bias-based, since – as we noted above – the bias-based distorted perceptual judgements are not within the circumstances that provide a reasonable basis for belief. This is because, as we saw earlier, such distorted perceptions violate an uncontroversial epistemic norm. On this construction of the reasonable person, individuals such as B are *unreasonable*, and the claim of self-defence is unavailable to them.

3.3 The Reasonable Person as the Non-culpable Person

An alternative reading of the reasonable person is as the person who may be mistaken or inaccurate, but non-culpably so. For example, Kelman offers a picture of the reasonable person as 'not blameworthy'; such that even if her beliefs are mistaken, they are not due to missteps that result from any fault on the part of the agent.[46] Likewise, Baron suggests that 'the fact that the belief is unreasonable should matter only if the individual is culpable for the belief'.[47] The idea is that even unreasonable beliefs should not be taken to reflect badly on the agent – and should not be beliefs she is held accountable for – if she is not culpable

[43] Baron (n 35) 27.
[44] *People v Goetz* (n 25).
[45] ibid 114.
[46] Kelman (n 39) 801.
[47] Baron (n 35) 26-7.

for arriving at those unreasonable beliefs. This adds an additional layer of evaluation into the reasonable person standard, and one that appears well-motivated when we consider cases such as the implicitly-biased individual. This view may seem to present a natural way of making sense of the individual, such as B, who believes that threat is imminent and lethal force is required, but believes this on the basis of distorted perceptual judgements; such an individual may suffer such distortions through no fault of her own.[48] This reading of the reasonable person standard makes the defence available to B, since it permits the conclusion that a person is reasonable if her beliefs, though unreasonable, are not culpably held. What is interesting about this construal is that, in fact, the reasonable belief requirement of US law drops out. The focus shifts to epistemic culpability, rather than whether the belief is reasonable.

We have, then, three versions of the reasonable person standard: the 'ordinary person standard', according to which the bias-based belief is reasonable; the 'reasonable basis for belief standard' according to which the bias-based belief is unreasonable; and the 'non-culpable standard', according to which what matters is not so much the unreasonableness of the belief, but whether the agent is culpable for holding it. In the following sub-section, we consider the ramifications of applying the latter two standards to bias-based belief. Since the 'ordinary person' standard is widely regarded as problematic, we set this aside.

4. Evaluating the Reasonable Person Standards

Our task in this section is to evaluate the relative merits and challenges that may face each of these formulations of the reasonable person standard. To recap, one option is the reading according to which bias-based beliefs are not reasonable (the reasonable basis for belief standard), so the claim of self-defence is unavailable to individuals such as B; another is the reading according to which, whilst bias-based beliefs are unreasonable, the agent is not culpable for having them and, therefore, is not unreasonable. Under the latter standard B's claim of self-defence may succeed. One way of proceeding with the comparative evaluation of these readings may be to consult our intuitions about reasonableness. But such intuitions are likely differing and flimsy, as evidenced by the inconsistencies characterising the literature and the case law. Instead, we can proceed by considering the wider ramifications of a legal system that endorses one or the other of these standards. It is at this point that Dan-Cohen's heuristic device, introduced earlier, comes to our aid.[49]

4.1 Reasonable Basis for Belief as a Decision-rule

A reasonable person standard that requires that individuals have a reasonable basis for belief is not met by defendants with bias-based beliefs; officials (judges instructing jurors, jurors themselves) would thus be guided by a standard designed to exclude from the

[48] Of course, if the individual does not disavow the biases and/or intentionally cultivates them, we may consider her culpable for harbouring them.

[49] Dan-Cohen (n 32).

realm of reasonableness the beliefs in the imminence of a threat and the necessity of force, where those beliefs are based on biased perceptual judgements – as in the case of B. In this respect, the standard tells officials to reject as unreasonable beliefs that are based in biases. This looks like a defensible standard, insofar as it asks officials to construe as unreasonable beliefs that have the hallmarks of defective, irrational or repugnant cognitions.

However, in considering such biased-based beliefs as unreasonable, and therefore excluding that individuals harbouring them can avail themselves of the claim of self-defence, this decision-rule may face various objections, each rooted in considerations of fairness. Considerations of fairness have been given (albeit brief) treatment by Lee in her discussion of implicit bias and self-defence: she considers the possibility that it is unfair to hold an individual liable 'for acting on a sincere belief that he was about to be killed … even if his beliefs stemmed from racially biased assumptions'.[50] One might support this line of thought, she remarks, by appealing to Garvey's claim that such a belief 'is one that only a saint or a fool would ignore'.[51] Garvey's point is that the state should not demand that anyone who genuinely believes that she is in great danger fails to heed that belief. Lee rejects this argument, on the assumption that allowing a claim to self-defence based on *any* sincere belief, irrespective of its reasonableness, is deeply problematic – we defer this issue to section 5, where we discuss English law. Lee also considers Garvey's suggestion that it would be illiberal for the state to punish citizens for having state-disapproved beliefs – e.g. racist beliefs – and that, therefore, the state cannot deny citizens a defence grounded upon such beliefs. However, Lee remarks that Garvey's position is not supported by the case law of the US Supreme Court.[52]

Whatever the merits of these responses to Garvey's sweeping claim that *any* sincere belief – even if explicitly grounded in racist stereotypes – should be a ground for self-defence, Lee's remarks do little to help us evaluate the fairness or otherwise of not accepting a claim of self-defence grounded in bias-based beliefs. Her remarks do not target the particular features of such beliefs which seem to make the charge of unfairness particularly apt. These features are as follows.

First, perceptual judgements influenced by implicit bias are difficult to avoid. This is especially so if individuals lack any knowledge of such biases, or have never considered that perceptual judgements themselves might be distorted by aspects of their cognition of which they are unaware.[53] Second, even if individuals are cognisant of such dispositions to bias, it is unclear that conditions of 'fair avoidability' are met, since bias mitigation

---

[50] Lee (n 30) 1604.

[51] Garvey (n 24) 126. Garvey's reasoning is that it may be problematic to deny the defence on the basis of bias-based beliefs since denying the defence to individuals such as B essentially makes individuals liable to punishment for certain attitudes they hold (e.g. implicit racial biases). Garvey worries that it is highly illiberal for the state to punish individuals on such a basis; or to hold them liable to punishment for failing to take sufficient steps to rid themselves from such attitudes. Note, though, that the criminal justice system often takes into account the attitudes of individuals both as evidence of guilt (see bad character evidence) and as constitutive of guilt (negligence).

[52] Lee (n 30) 1606.

[53] Cf. Washington and Kelly (n 23).

methods are not yet reliably successful. [54] Even an individual who took considerable measures to rid herself of biases could not be guaranteed success. Of those interventions that are successful, many are short-lived, with little evidence of any intervention being successful in the long term. As such, it may seem that individuals who use force due to bias-based beliefs may be unable to avoid any such erroneous beliefs, even when putting diligent effort into this task. A third and pressing line of concern with denying an individual the defence on the basis of biased beliefs is the extent to which this places liability upon an individual for what is effectively a collective failing. The concern here is that current understandings in social psychology attribute the causes of implicit biases to broader social and structural problems – prevalent stereotypes and inequalities that we may, as individuals, disavow. Insofar as this is the case, there is something problematic about holding an individual responsible – and accordingly, liable to punishment for intentional killing – on the basis of what is, essentially, a collective failure.

So, even if we were to agree with Lee in denying that any honest belief could per se ground a claim to self-defence, there are features of bias-based beliefs that make it particularly unfair to deny their reasonableness for the purposes of self-defence.

4.2 Reasonable Basis for Belief as a Rule of Conduct

Consider a conduct rule that demands that people ensure a reasonable basis for the beliefs that are relevant to self-defence. We are bound to conclude that it would be a violation of such a rule for someone to act in self-defence on the basis of bias-based beliefs. As a result, she is not permitted to use force. Individuals should conduct themselves with caution in instances in which there is reason to believe that bias may guide perception, belief, and – potentially lethal – action. Imposing a requirement to ensure a reasonable basis for belief resonates with the option favoured by Kelman – who does not, however, deal explicitly with implicit bias. He writes that in setting such a rule, 'we ask those who make at least partly race-based judgement of a person's violent intentions to use "alternative screening devices" … [B]asically, we ask them to wait until an actor makes his violent intentions clearer'.[55] The costs of doing so, he argues, are likely less than the costs of accepting race-based beliefs as reasonable for the purposes of self-defence. In any case, the conduct rule at issue would send a clear signal that force on the basis of bias-based beliefs is impermissible. It is valuable that the legal system conveys such strong anti-racist norms.

However, one might have doubts about the efficacy of such a rule of conduct in informing action, in particular when implicit biases are at play. First, the situations in which individuals deploy self-defence are most likely not situations in which careful deliberation is also deployed.[56] Second, as noted above, in the absence of knowledge about

[54] Lai (n 3).

[55] Kelman (n 39) 816.

[56] Cf. R Restak, 'The Fiction of the Reasonable Man' *The Washington Post* (Washington, 17 May 1987): https://www.washingtonpost.com/archive/opinions/1987/05/17/the-law-the-fiction-of-the-reasonable-man/15dea8f3-521a-48d0-aba8-9e361774450e/, making the rather strong claims that 'there are no reasonable people under conditions in which death or severe bodily harm are believed imminent' and that '[t]o expect

implicit bias, the efficacy of such a conduct rule is further reduced. Unless individuals are aware that their perceptions may be distorted by racial bias, an exhortation to wait until violent intentions are made clear may fail to receive uptake. After all, the bias distorts precisely those perceptions which represent violent intentions. Absent more widespread knowledge of implicit bias, then, concerns of fair avoidability arise. Moreover, even if the individual has knowledge of being biased, the automatic operation of the bias makes it extremely difficult for her to distinguish between biased and unbiased perceptions and to act only based on the latter. Perhaps an individual may train herself to distinguish those perceptions based in bias from those, which are undistorted; but the current state of knowledge and research does not reliably afford us such training techniques.[57]

4.3 Summary so far

Deployment of the 'reasonable basis for belief' interpretation of the reasonable person standard yields the conclusion that individuals who have bias-based beliefs in imminence and necessity are not reasonable, and therefore cannot avail themselves of the claim of self-defence. We have seen that various considerations of fairness arise in relation to the deployment of such an interpretation *qua* decision rule. And, *qua* conduct rule, whilst clear directives would be given by the law about the importance of avoiding race-based beliefs, current knowledge about implicit bias means that individuals are not well placed to guide their conduct in accordance with these directives. Let us now consider the alternative interpretation of the reasonable person standard, which permits mistaken and even unreasonable beliefs, so long as they are non-culpably held.

4.4 Non-culpable Unreasonable Belief as a Decision-rule

Under this interpretation of the reasonable person standard, the bias-based belief may well be unreasonable without undermining the reasonableness of the belief-holder. What matters, for the assessment of reasonableness of the belief-holder, is not the unreasonableness of the belief *per se*, but whether the individual is culpable for the unreasonable belief. On one common line of thought, to the extent that biases are pervasive, contingent upon social environment, not widely known of, and difficult to expunge from our cognitions, the distorted perceptual judgements and attendant beliefs are not culpable. The deployment of this reading of the standard as a decision rule would mean that the claim of self-defence would be available to individuals such as B.[58] The considerations of fairness aired above may speak in favour of this decision rule: even if the

---

reasonable behaviour in the face of perceived threat, terror and rage is itself a most unreasonable expectation'. Depending on how one construes the reasonable person standard, she may agree or disagree with either or both of Restak's claims. In any case, even if it were true that as a standard of conduct reasonableness could not serve as a deliberative guide in such contexts, this is not to say the reasonable person standard has no role, e.g. in providing the contents of a decision rule to criminal justice officials and in setting wider social expectations.

[57] Lai (n 3).

[58] Note that thus construed, it would make more sense for the claim of self-defence to function as an excuse, given that the relevant beliefs would be unreasonable. We address this option in section 6.

beliefs are unreasonable, the individual could not avoid the distorted judgement – at least, not without extraordinary measures – and wider patterns of social inequality or cultural stereotype are implicated in the individual's biases. The circumstances of the person are such that any of us may be similarly disposed to believe unreasonably. This construction of the standard instructs officials not to punish if the individual is guilty of no greater fault than most of the rest of us – namely, harbouring biases we disavow.

4.5 Non-culpable Unreasonable Belief as a Rule of Conduct

Recall that the claim of self-defence does double duty as a decision rule and a conduct rule. The defence serves as a justification, and so also conveys to the general public messages about how it is permissible to act.[59] If the claim of self-defence is successful in the case of bias-based belief, what conduct rule is thereby transmitted? The normative message conveyed would be that it is permissible to (attempt to) inflict harm or intentionally kill on the basis of bias-based beliefs. This seems deeply troubling for obvious reasons. First, it contains and conveys a disrespectful normative message. The message is that individuals may permissibly use potentially lethal force based on distorted perceptions that encode racist stereotypes. This devalues the lives of black citizens, allowing the reliance on distorted perceptions to outweigh their right to be protected by the state. Second, in maintaining that it is permissible to use force on the basis of bias-based beliefs, the state sanctions the deployment of racist stereotypes, permitting reliance on mistaken associations between black people and weapons to govern behaviour. Third, in sanctioning these stereotypes, the state plays a role in perpetuating and entrenching the very biases that distort cognition in the ways we have described. Fourth, the consequences of the pervasive knowledge of such a message may have an impact on the extent to which black citizens feel safe in making use of their freedoms: Kelman puts this in terms of being 'stigmatized, excluded from participation in generally available activities … [and] subjected to the demeaning supposition that others know a lot about them when who they truly are as individuals is wholly misassessed'.[60]

One might observe that the role of such a conduct rule is limited – indeed that rule may be utterly ineffective in guiding conduct – because the contexts in which it is deployed, contexts where threat appears imminent, are not ones in which deliberative thought and reflection on the directives of the law are generally gone in for. This does nothing to deflect the above worries. The conduct rule is troubling not because individuals in self-defence scenarios will deliberatively govern their conduct accordingly, but because

[59] Note that the concerns raised in this section could not be avoided even if the defence was not explicitly intended as a conduct rule, since conduct rules may be transmitted as unintentional side-effects of decision rules (cf. Dan-Cohen (n 32)). Moreover, if the non-culpable unreasonableness standard were understood as transforming self-defence into an excuse, rather than a justification, these concerns would not go away. True, the state would not say that it is permissible to act on such bias-based beliefs, but merely that doing so is excusable. But this would still convey – albeit perhaps with a lesser force – the problematic messages outlined below. More on this in section 6 below. Finally, note that on any rendering of the standard according to which bias-based beliefs are *reasonable* these concerns will arise with even greater force.

[60] Kelman (n 39) 816.

of the evaluative presuppositions of such a rule: at its starkest, that black lives don't matter, that racist stereotypes are a legitimate basis for action.

Nor are these worries mitigated by noting that the reasonable person standard exemplifies a strategy of 'selective transmission'. According to Dan-Cohen, strategies of selective transmission are deployed – intentionally or otherwise – to send different messages, respectively, to decision-makers and the general public (to whom rules of conduct apply).[61] Vagueness is one such strategy. Leaving a standard imprecisely specified may be one of the methods by which decision rules are insulated from the general public so as to avoid transmitting conduct rules which convey problematic normative messages.[62] The imprecision of a standard such as the reasonable person standard means that the law can withhold from explicitly committing to the claim that bias-based beliefs are a legitimate basis for the use of lethal force. This 'vagueness' may serve to cloak some of the more problematic aspects of a decision rule for determining reasonableness, and so may mitigate the problematic normative messages otherwise conveyed to the public.

Perhaps selective transmission is a possibility under certain circumstances, but it does not seem realistic in these cases. As a matter of empirical fact, great media attention is paid (and rightly so) to the outcomes of trials such as those of Goetz and those like our imaginary B. This is precisely because of concerns about racism in society and in criminal justice.[63] In any case, even if it were possible to deploy effective strategies of selective transmission, it seems to us that there is something sinister in the extreme about the law embedding an interpretation of a standard that sanctions racism, whilst concealing or attempting to conceal this from the general public. That the law deploys a legal construct which embeds racist messages, irrespective of whether they are heard loud and clear, is itself objectionable. Even if in general selective transmission strategies may be reconciled with rule of law requirements of clarity and publicity,[64] in this instance selective transmission is especially problematic.

4.6 Summary so far

This interpretation of the reasonable person standard allows the claim of self-defence to succeed in the case of bias-based beliefs, a result that appears to be supported by considerations of fairness towards the defendant. However, when we heed the implications that this rendering of the defence has in terms of rules of conduct (either

[61] Dan-Cohen (n 32) 635. Here is Dan-Cohen's example of a case in which selective transmission is useful (Dan-Cohen (n 32) 646): the conduct message 'ignorance of the law is no excuse' is well known and serves a useful function in setting a certain standard for conduct. But in practice, the decision rules deployed in determining whether conduct carried out in ignorance of the law is in fact punishable permit many exceptions. He argues that it is useful if these exceptions are not transmitted to the public, but remain embedded in the case law and in the legal scholarship surrounding it.

[62] As an example of how vagueness operates as a strategy of selective transmission, Dan-Cohen discusses the defence of duress. See Dan-Cohen (n 32) 639-40.

[63] See Caroline Light, *Stand Your Ground: A History of America's Love Affair with Self-Defense* (Beacon Press 2017) for work on the selective use of 'Stand Your Ground' laws in ways that favour white males.

[64] Dan-Cohen (n 32) 665-77.

explicitly, or as implied side-effects of the decision rule), we see that this construal of the reasonable person standard is deeply problematic and implicates the law in various racist evaluative stances.

The foregoing discussion, then, brings to light some deep tensions on either reading of the reasonable person standard. In short, a reading of the standard that denies self-defence – e.g. based on the consideration that implicit biases are not a reasonable basis for beliefs – faces deep concerns about fairness, at least given the current understanding of the cognitive phenomena at issue. On the other hand, a reading of the standard that permits the defence – based on the consideration that the relevant beliefs are non-culpably unreasonable – generates deeply problematic normative messages which entrench stereotypes and devalue the lives of black citizens. If the criminal law's reasonable person standard accommodates bias, racist normative messages are embedded in the law. If the criminal laws person is not biased, considerations of fairness arise in its treatment of citizens who, almost unavoidably in this historical moment, are so.

5. English Law and the Genuine Belief Rule

All of the above discussion is framed within the context of US law and focuses on the *reasonableness* or otherwise of the beliefs that a threat is imminent, and that potentially force is required to deflect the threat. One might think that these concerns are avoided in English law, according to which:

> The question whether the degree of force used by D was reasonable in the circumstances is to be decided by reference to the circumstances as D believed them to be (Criminal Justice and Immigration Act 2008, s.76(3)).[65]

Crucially, the individuals' beliefs (that a threat is imminent and force is required) do not have to be reasonable, but just genuinely held (the subjective element of the defence).[66] The requirement of reasonableness applies only to the degree of force used given the individual's subjective apprehension of the situation (the objective element).

English law, then, focuses on what the individual genuinely believed, and asks what force would be reasonable given that belief. In order to establish whether the use of force was reasonable, consideration is given to the fact that the individual, such as B, 'had only done what he honestly and instinctively thought was necessary'.[67] The issue that arises, given our present concern, is that we have reason to suppose that on at least some

[65] See also *R. v Gladstone Williams* (1984) 78 Cr App R 276.

[66] The European Court of Human Rights claimed that a genuine belief is not sufficient; the belief must also be held 'for good reason' (see *McCann v United Kingdom* (1996) 21 EHRR 97, para 200). Notably, the Court only dealt with cases of preventative force under Art. 2 of the European Convention on Human Rights (Right to life), thus cases involving the killing of an individual on the part of enforcement officers. In any case, notwithstanding that it had the opportunity to do so, the Court has not remarked on the incompatibility between the Convention and the English law on self-defence.

[67] Criminal Justice and Immigration Act 2008, s 76 (7)(b).

occasions, what individuals 'honestly and instinctively' believe to be necessary is the result of distorted perceptual judgements. An individual might believe that force is necessary on the basis of the belief that there is a threat, and that belief – in the presence of a weapon, or in the degree of hostility manifested – may be based on distorted perceptual judgements.

5.1 Evaluating the 'Genuine Belief' Standard

Avoiding an assessment about the reasonableness of the belief does not resolve the tensions outlined in section 4. On the face of it, as a decision rule, the English standard directs officials to maintain that even unreasonable but genuinely held beliefs are compatible with the availability of the defence. Accordingly, the genuine belief standard renders acting on bias-based beliefs justified, so long as the force used was reasonable, given such beliefs. So understood, the genuine belief standard would appear to face all the problematic ramifications outlined above: conveying disrespect; sanctioning cognitions that deploy racist stereotypes; entrenching those stereotypes, and the demeaning consequences of stigmatisation for those individuals targeted by them. Yet, avoiding these consequences would require refinement of the genuine belief standard. One might either creatively interpret the standard so as to deny that such bias-based beliefs could be 'honestly and instinctively' held; there is something inherently *dishonest* about such biased beliefs, one might say. Alternatively, one could simply hold that such bias-based beliefs are insufficient for the purposes of self-defence; even honest belief, when based in bias, cannot justify defensive force. Either of these refinements, though, faces the fairness concerns for the defendant, as raised above. So, the very same tension discussed earlier seems to play out in the context of the genuine, rather than reasonable, belief rule. However, there are additional complexities to the way that English law deals with self-defence that are worth addressing.

5.2 The Re-emergence of the Reasonable Person

Whilst the focus of the English defence is on *genuine* belief, the notion of reasonableness re-emerges in the instructions regarding how one might ascertain genuineness of belief. Section 76(4) Criminal Justice and Immigration Act 2008 indicates that:

> If D [the defendant] claims to have held a particular belief as regards the existence of any circumstances –
>
> (a) the reasonableness or otherwise of that belief is relevant to the question whether D genuinely held it.[68]

Thus, officials are instructed that certain inferences may be made: from normative

[68] Cf. CPS Legal Guidance, 'Self-Defence and the Prevention of Crime: S76 of Criminal Justice and Immigration Act 2008' http://www.cps.gov.uk/legal/s_to_u/self_defence/#rachel which also emphasise the evidential role of reasonableness. Crucially, these guidelines also suggest that 'the more unreasonable the belief, the less likely it is that the court will accept it was honestly held'.

standards to the existence of cognitive states.[69] That a belief is reasonable may be evidence that the individual held it; that it is unreasonable may be evidence that the belief was not honestly held. The reasonable person standard re-emerges as a decision rule, in particular, with an evidential role. That a reasonable person may have believed *p* is said to provide evidential support for the fact that the defendant did genuinely so believe. This raises again the issue as to what a reasonable person might believe – in particular, whether she may form bias-based beliefs.

Given the pervasive disposition to bias-based beliefs, this evidential directive seems warranted only if the model of a reasonable person is construed as incorporating the sorts of implicit biases outlined in section 1. But as we have seen, to endorse such a model is to face the objections raised in section 4. This is so, even if provision 4(a) above is explicitly indicated as a decision-rule: as we observed earlier conduct rules may be generated as a side-effect, conveying the normative message that it is reasonable to hold and act on bias-based beliefs. Admittedly, given that this decision rule would be evidential rather than substantive in nature – in particular, it would not identify a fact that is relevant for criminal responsibility, but only a fact that is relevant to prove one such fact – it is plausible to argue that the rule would be less powerful in sending the negative messages discussed earlier, as it would play a less visible role in adjudication. Accordingly, this evidential role may be an effective method of 'selective transmission', by which the problematic messages embedded in the evidential directive are at least in part prevented from conveying problematic messages about permissible conduct. However, as we argued earlier, a legal construct which embeds racist assumptions is itself objectionable, and covering it up is no remedy.

On the other hand, a model of the reasonable person which does not accommodate within the circumstances of the reasonable person such biases undermines the evidential role that reasonableness should play according to s 76(4). Under this reading, the fact that a belief is reasonable gives little evidential support to the claim that it was genuinely held, since we know that dispositions to bias pervasively influence our cognition. A model of the reasonable person that excludes implicit bias, then, considerably weakens the epistemic warrant for this particular decision rule.

6. A Palliative Solution

The tension we have identified is between fairness towards the defendant and the sanctioning of racist stereotypes. This arises most starkly in judicial systems in which the only options facing court officials are to convict an individual for an intentional infliction of harm (or attempt thereby), or to accept the justification of self-defence, and thus acquit. As Lee notes, a middle ground may be to resort to a partial defence of *imperfect self-defence*

[69] Cf. *Palmer v The Queen* [1971] AC 814, para 832, defending the reverse of this inference: 'If a jury thought that in a moment of unexpected anguish a person attacked had only done what he honestly and instinctively thought was necessary that would be most potent evidence that only reasonable defensive action had been taken'.

in a case of genuine but unreasonable belief.[70] This avenue is followed in several US jurisdictions. Such a defence may serve to reduce the charge – for example, from murder to voluntary manslaughter.[71] This partial defence could provide a model for the treatment of self-defence claims in cases of non-culpably-unreasonable belief. It could then be available to individuals who use force due to bias-based beliefs.

Lee's proposal is instructive because it allows us to identify one variable that is relevant to the tensions articulated above: partial vs. full defence. The merit of a partial defence, in our case of bias-based belief, is that it mitigates the problematic messages that a full defence may send to the public, whether directly or as a side-effect. However, the partial defence still apportions punishment – perhaps severe – to the defendant, so that concerns of fairness in the face of non-culpable mistake remain. Grave concerns persist about punishing individuals – albeit with a more lenient sentence – for intentional harm perpetrated on the basis of cognitions they not only repudiate, but inherited from a social context shaped by a state whose institutional structures and dynamics have demonstrated a disregard for racial equality.

There is a second independent variable that is not made explicit in Lee's discussion of the imperfect defence: excuse vs. justification.[72] As observed at the start, self-defence may function as a justification; that is, it renders the conduct permissible. If the defence functions as an excuse, instead, it does not have this implication; the conduct is treated as wrongful notwithstanding the defence. The merit of excusing over justifying in the case of bias-based belief, then, is that the defence sends a qualitatively different message to the public. It conveys the message that individuals are *not* permitted to act on racial biases, although they are excused for doing so. Note, though, that from a perspective of dissatisfaction with a society infected by racism and with the role of criminal justice within it, this message remains inadequate. It communicates that intentionally killing someone due to racist stereotypes is excusable. Such a message would legitimately be met with outrage.[73] This is especially the case where the criminal justice system itself is deeply

---

[70] Lee (n 30). The main proposals that Lee considers are ways of 'making race salient' in the criminal trial, since empirical evidence suggests that doing so is an effective means to reducing racism (1586-1600). For example, mock jurors encouraged to consider race directly across various scenarios avoided judgements that expressed racial bias, compared to those for whom race was not made salient. Whilst there is much of interest in here for criminal justice systems and theorists to consider, Lee's proposal to deploy such strategies in self-defence cases is based on the assumption that bias-based beliefs are obviously unreasonable and, therefore, that they could not ground self-defence. This stance, however, does not address the distinctive considerations of unfairness that we raised in section 4.

[71] Lee (n 30) 1603-1604. A somewhat similar approach is taken in the Model Penal Code, notwithstanding the absence of any mention of reasonableness. Cf. the Model Penal Code, s 3.09(2), stating that if the belief is recklessly or negligently mistaken and if recklessness and negligence suffice to establish culpability for the relevant crime, the defendant should be convicted as if she acted recklessly or negligently. This looks like a legal fiction, given that in cases such as B's the agent acts intentionally.

[72] For a discussion of this variable in the context of self-defence, see Alan Norrie, *Crime, Reason and History: A Critical Introduction to Criminal Law* (4th edn, Cambridge University Press 2014) 292-4 and George Fletcher, *Basic Concepts of Criminal Law* (Oxford, Oxford University Press 1998) 130-8, 158-63.

[73] One might think that this message could be nullified by accompanying the application of the defence with

implicated in entrenching racial inequality.[74]

Any of the options available within this two-variable framework, then (consider, for instance, a partial excuse), may lessen but does not resolve the tensions that we have articulated. The best that this framework can offer is a schema of palliative responses to the problems that implicit racial biases generate in the case of self-defence.[75]

7. Concluding Remarks

We set out with a hypothetical – but all too likely – scenario where implicit racial biases are implicated in perceptual judgements, on the basis of which beliefs about the imminence and gravity of a threat are formed and acted upon. Is the criminal law's person someone who would form such bias-based beliefs? We have argued that there are costs on either way of settling this question. If the criminal law's reasonable person is not susceptible to implicit biases, a distinctive question of fairness arises. The fact is that such biases may influence all of us despite our best efforts. Under these circumstances it is unfair to punish individuals for acting in accordance with bias-based beliefs. Yet if the criminal law's reasonable person is susceptible to such implicit biases, the system embeds racist stereotypes. This communicates problematic normative messages about the legitimacy of deploying racial stereotypes, and devalues black lives.

It is important to note that this tension is not atemporal and irresolvable. Here the challenges to an adequate model of the criminal law's person are distinctive. They differ from those arising from lines of argument to the effect that all agents lack the requisite form of agency to ever be criminally responsible – because of some metaphysical thesis of determinism, or because of neurological features that show our agency to be mechanistic and beyond our control. Rather – and by way of final diagnostic remarks – we propose that this tension arises from the following two features of our relationship to criminal law. The first is contingency: actual human agency, it's functioning, and what can reasonably be expected of us all, is contingent upon our social context. We are susceptible to bias because our cognitions are shaped by a racist history and environment; but our cognisance of these

---

an explicit pronouncement, disavowing such stereotypes. But it appears manifestly inadequate to disavow racial stereotypes at the same time as embedding a concept that deploys them.

[74] Michelle Alexander, *The New Jim Crow: Mass Incarceration in the Age of Color Blindness* (New York, The New Press).

[75] Someone may raise issues concerning the implementation of any proposal that would require ascertaining the role of implicit bias in the defendant's action. For instance, she may contend that it would be infeasible to provide evidence not only that the defendant was implicitly biased, but also that these biases influenced her beliefs at the time of acting. However, one can provide evidence of the presence of a bias – as has been done millions of times over online – with use of the IAT (or commensurate measure). See Project Implicit: https://implicit.harvard.edu/implicit/. Also, one proxy indicator of the bias' causal role in action could be the strength of the association measured. Finally, it seems to us that the problems raised by this objection are not qualitatively different from problems that beset other defences with a long-standing legal pedigree: consider defences of insanity, loss of control, diminished responsibility and intoxication. It is arguable that the expert evidence that would be available to establish the facts relevant to the defence discussed here is on no less solid footing than the (expert or non-expert) evidence often relied upon in order to establish these other defences.

psychological and social phenomena and our strategies in addressing them are not yet sufficiently developed that we may reasonably be expected to avoid such defective agency. However, the contingent facts of our racist social context make all the more pressing the need to fashion a criminal justice system that does not perpetuate further stigmatising stereotypes. Of course, the criminal law could not and should not be the only tool through which to expunge implicit biases or secure racial equality; but, at least, it should not embed racist presuppositions, fuel implicit biases, and perpetuate racist stigmatisation.

The second feature that gives rise to the tension we have identified is the dual role that we ask the notion of the person in criminal law to perform. This notion is sometimes invoked as a normative ideal, setting standards to which we should strive in our conduct. Sometimes, however, the notion aims to accommodate the many ways in which, perhaps through no fault of our own, we may be defective.[76] Our discussion points to the somewhat pessimistic conclusion that where our agency is embedded in socially unjust relations, the normative ideal is a rather distant, perhaps unachievable one.

References

Alexander M, *The New Jim Crow. Mass Incarceration in the Age of Color Blindness* (New York, The New Press 2012).

Amodio D and Devine P, 'Stereotyping And Evaluation In Implicit Race Bias: Evidence For Independent Constructs And Unique Effects On Behaviour' (2006) 91 *Journal of personality and Social Psychology* 91, 652.

Armour JD, 'Race Ipsa Loquitur: Of Reasonable Racists, Intelligent Bayesians, and Involuntary Negrophobes' (1994) 46 *Stanford Law Review* 781.

Ashworth A and Horder J, *Principles of Criminal Law* (7th edn Oxford University Press 2013).

Baron M, 'The Standard of the Reasonable Person in Criminal Law' in Duff, Farmer, Marshall, Renzo, and Tadros (eds), *The Structures of Criminal Law* (Oxford University Press 2011).

[76] This is most evident when the criminal law's person is construed based on the (problematic) notion of normality and on the traits of the agent. A clear example of this is the English law on loss of control. See, in particular, s 54(1)(c) of the Coroners and Justice Act 2009. For a critical assessment of the tension between the reasonable person standard as normative ideal and its descriptive/exculpatory content see Moran (n 39) 301-7.

Blank RM, Dabady M and Citro CF (eds), *Measuring Racial Discrimination* (Washington DC, The National Academies Press 2004).

Dan-Cohen M, 'Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law' (1984) 97 *Harvard Law Review* 625.

Devine, PG, Ashby Plant E, Amodio DM, Harmon-Jones E, and Vance SL, 'The Regulation Of Explicit And Implicit Race Bias: The Role Of Motivations To Respond Without Prejudice' (2002) 82 *Journal of Personality and Social Psychology* 835.

Devine PG, Forscher PS, Austin AJ and Cox WT, 'Long-term Reduction in Implicit Race Bias: A Prejudice Habit-breaking Intervention' (2012) 48 *Journal of Experimental Social Psychology* 1267.

Dovidio JF and Gaertner SL, 'Aversive Racism and Selection Decisions: 1989 and 1999' (2000) 11 *Psychological Science* 315.

Dovidio JF, Gaertner SL, Kawakami K and Hodson G, 'Why can't we just get along? Interpersonal biases and interracial distrust' (2002) 8 *Cultural Diversity and Ethnic Minority Psychology* 88.

Duncan B, 'Differential Social Perceptions and the Attribution of Intergroup Violence: Testing the Lower Limits of Stereotyping of Blacks' (1976) 34 *Journal of Personality and Social Psychology* 590.

Eberhardt J, 'Seeing Black: Race, Crime and Visual Processing' (2004) 87 *Journal of Personality and Social Psychology* 876.

Fletcher G, *Basic Concepts of Criminal Law* (Oxford University Press 1998).

Fletcher G, *A Crime of Self-Defense: Bernhard Goetz and the Law on Trial* (University of Chicago Press 1988).

Gardner J, 'The Many Faces of the Reasonable Person' (2015) 131 *Law Quarterly Review* 563.

Garvey SP, 'Self-Defense and the Mistaken Racist' (2008) 11 *New Criminal Law Review: In International and Interdisciplinary Journal* 119.

Gendler TS, 'On the epistemic costs of implicit bias' (2011) 156 *Philosophical Studies* 33-63.

Glaser J and Knowles ED, 'Implicit Motivation to Control Prejudice' (2008) 44 *Journal of Experimental Social Psychology* 164.

Green AR, Carney DR, Palin DJ, Ngo LH, Raymond KL, Iezzoni, LI and Banaji MR, 'Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients' (2007) 22 *Journal of General Internal Medicine* 1231.

Haslanger S, 'Distinguished Lecture: Social Structure, Narrative and Explanation' (2015) 45 *Canadian Journal of Philosophy* 1.

Holroyd J, 'The Retributive Emotions: Passions And Pains Of Punishment' (2010) 39 *Philosophical Papers* 343.

Holroyd J, 'Responsibility for Implicit Bias' (2012) 43 *Journal of Social Philosophy* 274.

Holroyd J, 'Implicit Bias, Awareness And Imperfect Cognitions' (2014) 33 *Consciousness and Cognition* 511.

Holroyd J, 'What do we Want from a Model of Implicit Cognition?' (2016) 112 *Proceedings of the Aristotelian Society* 153.

Holroyd J and Kelly D, 'Implicit Bias, Character and Control' in Webber J and Masala A (eds), *From Personality to Virtue* (Oxford, Oxford University Press 2016).

Holroyd J and Puddifoot K, 'Implicit Bias and Prejudice' in Fricker M, Graham P, Henderson D, Pedersen N and Wyatt J (eds), *The Routledge Handbook of Social Epistemology* (Taylor and Francis 2019) (forthcoming).

Hugenberg K and Bodenhausen G, 'Facing prejudice implicit prejudice and the perception of facial threat' (2003) 14 *Psychological Science* 640-3.

Jost J, Rudman L, Blair I, Carney D, Dasgupta N, Glaser J and Hardin C, 'The Existence of Implicit Bias is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of 10 Studies that no Manager Should Ignore' [2009] *Research in Organisational Behaviour* 29.

Kelman M, 'Reasonable evidence of reasonableness' (1991) 17 *Critical Inquiry* 798.

Lacey N, 'Socializing the Subject of Criminal Law: Criminal Responsibility and the Purposes of Criminalization' (2016) 99 *Marquette Law Review* 541.

Lai, CK, Marini M, Lehr SA, Cerruti C, Shin JL, Joy-Gaba JA and Ho AK, 'Reducing

implicit racial preferences: I. A comparative investigation of 17 interventions' (2014) 143 *Journal of Experimental Psychology: General* 1765.

Lee C, 'Making race salient: Trayvon Martin and implicit bias in a not yet post-racial society' (2013) 91 North Carolina Law Review 1555.

Levy N, 'Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements' (2015) 49 *Nous* 800.

Light C, *Stand Your Ground: A History of America's Love Affair with Self-Defense* (Beacon Press 2017).

Madva A, 'Why Implicit Attitudes are (probably) Not Beliefs' (2016) 193 *Synthese* 2659.

Moran M, *Rethinking the Reasonable Person. An Egalitarian Reconstruction of the Objective Standard* (Oxford University Press 2003).

Norrie A, *Crime, Reason and History. A Critical Introduction to Criminal Law* (4th edn, Cambridge University Press 2014).

Nosek GB and Banaji M, 'The Implicit Association Test at Age 7: A Methodological and Conceptual Review' in Bargh J (ed), *Automatic Processes in Social Thinking and Behaviour* (Psychology Press 2007).

Payne KB, 'Weapon Bias Split-Second Decisions And Unintended Stereotyping' (2006) 15 *Current Directions in Psychological Science* 287.

Payne KB, Shimizu Y and Jacoby LL, 'Mental control and visual illusions: Toward explaining race-biased weapon misidentifications' (2005) 41 *Journal of Experimental Social Psychology* 36.

Puddifoot K, 'Dissolving the Ethical/Epistemic Dilemma over Implicit Bias' *Philosophical Explorations (*forthcoming).

Restak R, 'The Fiction of the 'Reasonable Man' *The Washington Post* (17 May 1987).

Robinson PH, *Criminal Law: Case Studies and Controversies* (Aspen Publishers 2005).

Saul J, 'Implicit Bias, Stereotype Threat and Women in Philosophy' in Hutchinson K and Jenkings F (eds), *Women in Philosophy: What Needs to Change?* (Oxford, Oxford University Press 2013).

Scaife R, Holroyd J, Stafford T and Bunge A, 'The Effects of Moral Interactions on Implicit Racial Bias' (MS).

Siegel S, 'Cognitive penetrability and perceptual justification' (2012) 46 *Noûs* 201.

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

Washington N and Kelly D, 'Who is Responsible for this?' in Brownstein and Saul (eds), *Implicit Bias and Philosophy* (Oxford University Press 2016).