Supplementary information: Additional details of experiments

Experiment 1

IATs:

Race IAT used in all following experiments:

Positive words: Good, Happy, Joy, Love, Pleasure.

Negative words: Agony, Bad, Evil, Hurt, Nasty.

Images of cropped faces taken from project implicit stimulus materials (http://www.projectimplicit.net/stimuli.html): Race faces from the ‘race attitude’ task.

Muslim IAT:

Positive words: Good, Happy, Joy, Love, Pleasure.

Negative words: Agony, Bad, Evil, Hurt, Nasty.

Typical Muslim names: Fatima, Mohammed, Saad, Yusra, Zahir.


Explicit questions:

A) In general how do you feel about black people?

7 point scale semantic differential (Cold-Warm/Suspicious-Trust/ing/ Positive-Negative/Friendly-Hostile)

B) In general how do you feel about white people?

7 point scale semantic differential (Cold-Warm/Suspicious-Trust/ing/ Positive-Negative/Friendly-Hostile)
C) In general how do you feel about Muslims?

7 point scale semantic differential (Cold-Warm/Suspicious-Trust/ Positive-Negative/Friendly-Hostile)

Likert scale questions (agree/disagree 7 point scale):

1. I believe racial integration is beneficial for this country

2. I am motivated by my values to avoid prejudice on the basis of racial identity

3. I would be happy to work with a student from another race in assessed group work

4. I believe religious diversity is beneficial for this country

5. I am motivated by my values to avoid prejudice on the basis of religious identity

6. I would be happy to work with a student from another religious group in assessed group work

7. It is morally acceptable to make judgements about people based on their race

8. I would feel guilty if I found that I behaved in a prejudiced way towards different racial groups

9. I would feel ashamed if my friends believed that I harboured racially prejudiced feelings

10. It is socially unacceptable to make judgements about people based on their race

11. I am less racially prejudiced than most people in my peer group

12. It is morally acceptable to make judgements about people based on their religious identity

13. I would feel guilty if I found that I behaved in a prejudiced way towards different religious groups
14. I would feel ashamed if my friends believed that I harboured prejudiced feelings towards different religious groups

15. It is socially unacceptable to make judgements about people based on their religion

16. I am less prejudiced towards other religious groups than most people in my peer group

**Experiment 2**

Blame condition script:

‘You have just taken the shooter bias test, which is intended to measure differences in attitudes towards racial groups that you might not explicitly endorse.

I'm afraid that the differences in your reaction times and shooting choices indicate you have negative implicit attitudes towards black people.

Morally speaking, we would hope people don't have these kinds of attitudes. People who have these kinds of attitudes tend to behave in discriminatory ways, even if it is so subtle that you don't notice it.

Overall, you are blameworthy for having these discriminatory attitudes and behaviours. As you probably know, it is morally unacceptable to have biased attitudes and behaviours.

It would be quite normal to feel guilty about this; and to think about how to change these attitudes, or your behaviours to bring them in line with moral expectations.

Later, in the debrief, we can talk more about techniques people have used to try to eliminate these bad attitudes. There'll also be the chance to ask any questions you may have.

Now that you've got the results of this part of the study, we'll give you a moment to reflect on that, and then move on to the next part of the study.’
Exculpation condition script:

'You have just taken the shooter bias test, which is intended to measure differences in attitudes towards racial groups that you might not explicitly endorse.

I'm afraid that the differences in your reaction times and shooting choices indicate that you have negative implicit attitudes towards black people.

Morally speaking, we would hope people don't have these kinds of attitudes, but it is actually entirely blameless. People who have these attitudes tend to behave in discriminatory ways, but it is so subtle you don't notice it, and you aren't culpable for doing so.

It would be quite easy to feel guilty about these attitudes and behaviours, but you shouldn't. In fact, it would be great if you were concerned to consider the various steps that can be taken to change these attitudes, or your behaviours.

Later, in the debrief, we can talk more about techniques people have used to try to eliminate these attitudes. There'll also be the chance to ask any questions you may have.

Now that you've got the results of this part of the study, we'll give you a moment to reflect on that, and then move on to the next part of the study.'

Explicit questions:

Questions A-B and 1-3, 7-11, from experiment 1.

**Experiment 2b**

Explicit questions:

1. What can you remember about taking part in the initial experiment last semester? (open)
2. After receiving the feedback in the initial experiment did you feel…? (1-7 scale)

a. Blamed  
b. Guilty  
c. Anxious  
d. Relaxed  
e. Upset  
f. Excused  
g. Self-conscious  
h. Angry  

3. Did you feel responsible for being prejudiced? (1-7 scale)  

4. Can you remember if you were blamed during your feedback on the shooter bias task? (I was blamed and told it would be normal to feel guilty/ I was not blamed and told not to feel guilty/I don’t remember which feedback I received).  

5. Do you think taking part in the initial experiment has made you less likely to make prejudiced judgments? (Yes/No)  

6. Before today have you thought about the initial experiment since taking part (1-7 scale)  

7. Have you done anything to try to ensure you treat all people equally since taking part in the initial experiment? (Yes/No)  

a) If yes what have you done (Open)  

8. Have you done any experiments on racial associations since taking part in the initial experiment? (Yes/No)  

a) If yes how many (open)  

9. How do you feel about having taken part in the initial experiment rather than getting credits from other psychology experiments instead? (1-7 scale: I’m glad I took part-I regret taking part).  

Experiment 2b additional analysis:
The figure above shows the correlation between IAT scores of participants in experiment 2 (time point 1) and the IAT scores of the same participants in experiment 2b (time point 2) which took place 6 months later. It is worth noting that this is not a clean test-retest analysis because experiment 2 involved a manipulation before taking the IAT. As such some of the difference may be due to any effect of that manipulation fading over time.

Note that IATs here are coded so that negative scores represent the stereotypical (prejudicial) associations.

<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Total Responses</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blamed</td>
<td>9</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>16</td>
<td>6</td>
<td>5</td>
<td>60</td>
<td>3.78</td>
</tr>
<tr>
<td>2</td>
<td>Guilty</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>18</td>
<td>17</td>
<td>11</td>
<td>60</td>
<td>5.12</td>
</tr>
<tr>
<td>3</td>
<td>Anxious</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>17</td>
<td>14</td>
<td>7</td>
<td>4</td>
<td>60</td>
<td>4.17</td>
</tr>
<tr>
<td>4</td>
<td>Relaxed</td>
<td>6</td>
<td>25</td>
<td>18</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td>2.68</td>
</tr>
<tr>
<td>5</td>
<td>Upset</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>6</td>
<td>21</td>
<td>7</td>
<td>7</td>
<td>60</td>
<td>4.40</td>
</tr>
<tr>
<td>6</td>
<td>Excused</td>
<td>13</td>
<td>18</td>
<td>12</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td>2.63</td>
</tr>
<tr>
<td>7</td>
<td>Self-conscious</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>13</td>
<td>18</td>
<td>9</td>
<td>7</td>
<td>60</td>
<td>4.55</td>
</tr>
<tr>
<td>8</td>
<td>Angry</td>
<td>15</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>60</td>
<td>3.30</td>
</tr>
</tbody>
</table>

The table above shows how the participants in experiment 2b remember their emotional state after taking part in experiment 2. These ratings were elicited by asking: “After receiving the feedback in the initial experiment did you feel…?” and then presenting each of the emotions on the table using a 7 point Likert scale.
Experiment 3

Explicit questions:

Questions A-B, and 1-3, 7-11, from experiment 1.

Additional explicit questions following the race IAT:

Sorting when the categories are black with good and white with bad was…

(1) A lot easier, (2) moderately easier, (3) slightly easier, (4) the same, (5) slightly harder, (6) moderately harder, (7) a lot harder…

Than sorting when the categories were white with good and black with bad.

Did this surprise you? (Yes-No 7 point scale)

Has taking part in the experiment lead you to better understand your own racial attitudes (Yes-No 7 point scale)

Do you intend to try to change your future behaviour as a result of your experience in this experiment? (Yes-No 7 point scale)

Emotion questions

Considering your experience of the experiment so far to what extent do you feel:

Humility questions


Additional exploratory analysis from experiment 3

Our pre-registered inclusion criteria specified having English as a first language (i.e. without consideration of where a person was from or their ethnicity). However, this exclusion criteria was not effective in allowing us to focus on those who grew up in a predominantly English speaking culture nor did it narrow the sample to a particular ethnic group. As a number of our sample were non-white British native English speakers. If we re-analyse the results from experiment 3 and use the stronger exclusion criteria of only including the results of the white participants who were native English speakers this increases the effect size (to 0.28, 95% confidence interval -0.06, 0.62). Keeping these stricter exclusion criteria, if the blame conditions from experiments 2 and 3 are pooled and compared to the control conditions from experiment 3 then the difference is significant (blame mean -0.18, SD 0.59; control mean -0.41, SD 0.57; t(177)= 2.49 p=0.01, effect size: 0.38). Given that this is exploratory analysis we should not read much, if anything, into this finding but it does perhaps indicate that in future experiments it may be worth narrowing the focus to white native English speaking participants.

Behavioural Intentions:

Participants were asked: ‘Do you intend to try to change your future behaviour as a result of your experience in this experiment?’ This question provided evidence that the blame manipulation has a positive impact on people’s explicit intentions to change their future behaviour (mean 5.03, SD 1.77) when compared to the behavioural intentions of the control group (mean 3.48, SD 1.86), t(157)=5.38, p<0.000001. Effect size: 0.85. This is so despite the fact that the communication of blame was an uncomfortable and intense interaction producing a significant increase in anxiety (blamed mean 3.46, SD 1.72, control mean 2.77 SD 1.67; t(160)=2.60,p=0.01. Effect size: 0.41). Furthermore, the degree of intended behaviour change correlates positively with the degree to
which participants felt blamed $r(160)=0.22$, $p=0.0048$. It also correlates positively with the degree to which participants felt guilty $r(160)=0.17$, $p=0.03$.

The seating distance measure and voluntary time commitment measure yielded no significant difference between conditions. It is not possible to determine which of the various competing explanations is at work here. Possibilities include: the measure failing to detect an effect; the absence of any behavioural effect to detect; the possibility that the behaviours accessed by the measures did not correlate with the association measured on the IAT; the fact that a one-shot measure such as this may not reveal behavioural dispositions that are revealed in patterns over time.

Awareness:

Since awareness of when bias has been displayed is important for individuals’ ability to identify when remedial strategies are needed, we also included explicit questions asking participants to report awareness of their own implicit prejudice by indicating if they found it more or less difficult to associate black faces with positive words than white faces. The reported difficulty correlates negatively with participants’ IAT scores $r=0.401$, $n=162$, $p<0.000001$. This supports previous work which has shown that people are not necessarily unaware of their implicit biases and have found participants can accurately estimate their own degree of implicit bias, even without having taken an IAT (Hahn, Judd, Hirsh, & Blair, 2014). The basis on which our participants reported such awareness may be driven by their recent observation of their own behavioural responses on the IAT (Holroyd, 2015). Of course, this raises questions about the relevance of this finding to the possibility of awareness of implicitly biased behaviour outside the lab. In such non-laboratory scenarios individuals may not have been prompted to reflect on their responses, nor had the distinctive experience of taking an IAT upon which to reflect. It is worth noting, though, that there was nothing about the prompting of awareness to prevent this technique from being deployed outside the lab, nor to suggest that it was unduly 'leading' in garnering awareness. Whilst the question about awareness drew attention to individuals' responses on the IAT, the question itself did
not lead individuals towards any particular answer. The accuracy of responses cannot be explained without appeal to individuals' awareness. Second, whilst the experiences on which individuals reflected were specific to the experiment (response time to IAT) there are various other more familiar aspects of experience – feelings of discomfort or stress – in which implicit biases may manifest. These may provide the basis for such observational awareness outside the laboratory.

**Surprise:**

If individuals are able to be aware of implicit bias, we might expect their experience on the IATs not to reveal anything new to them. Accordingly, we incorporated a question about how surprised they were about their experiences of taking the IAT (in particular, their awareness of the comparative difficulty of sorting black paired with positive words and sorting white paired with positive words). There was a strong correlation between difficulty on stereotype-incongruent trials and surprise $r(160)=0.38$, $p<0.000001$. This indicates that the more difficulty participants experience when pairing black with good words the more surprised they are about this. There was also a negative correlation between IAT score and surprise $r(160)=-0.17$, $p=0.03$. This indicates that those with more anti-black prejudiced IAT scores were more surprised by their experience taking the IAT. Both these findings support the view that participants are gaining an insight into their biases from their behavioural experience rather than directly through introspection.

**Humility:**

An exploratory analysis of the characterological correlates of openness to change was undertaken. We used the Intellectual Humility Scale which is a 13-item scale developed by McElroy et al (2014). This scale features questions concerning whether participants love learning and whether they are comfortable with and take ownership of their intellectual limits. Intellectual humility did not predict IAT scores ($r(160)=0.06$, $p=0.47$). This indicates that those high in intellectual humility were no more or less biased than those lower in intellectual humility. There was also no significant correlation between IAT scores and intellectual humility when the blame condition is considered in
isolation ($r(77)=0.11$, $p=0.33$). This suggests those high in intellectual humility were not more influenced by moral feedback, as would have been indicated by a significant negative correlation. There was a negative correlation between intellectual humility and intention to change future behaviour $r(160)=-0.21$, $p=0.008$. This also suggests that intellectual humility does not indicate an increased receptiveness to feedback. This is surprising, given common sense assumptions about what intellectual humility consists in. Since the scale is in early stages of development, it may be that this surprising result points to the need for further refinement of the scale. However, this may be because those high in intellectual humility were more likely to already be aware of their implicit biases and thus less likely to change behavioural intentions as a result of taking part in the experiment. This interpretation is supported by the fact that intellectual humility correlates negatively with surprise regarding the level of difficulty experienced when paring black faces and positive words in the IAT ($r(160)=-0.20$, $p=0.012$). There were relationships between intellectual humility and how taking part in the experiment made participants feel — intellectual humility was associated with greater positive emotions and lesser negative emotions (see Table 1). For example, those high in intellectual humility felt less blamed and less guilty. These associations are consistent between conditions, with a tendency for stronger relationships between these emotions and intellectual humility in the blame condition.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>anxious</th>
<th>relaxed</th>
<th>upset</th>
<th>happy</th>
<th>angry</th>
<th>calm</th>
<th>blamed</th>
<th>guilty</th>
<th>embarrassed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>-0.29</td>
<td>0.22</td>
<td>-0.22</td>
<td>0.17</td>
<td>-0.15</td>
<td>0.21</td>
<td>-0.32</td>
<td>-0.25</td>
<td>-0.31</td>
</tr>
<tr>
<td>Blame</td>
<td>-0.29</td>
<td>0.35</td>
<td>-0.37</td>
<td>0.33</td>
<td>-0.21</td>
<td>0.27</td>
<td>-0.32</td>
<td>-0.38</td>
<td>-0.37</td>
</tr>
</tbody>
</table>

Table 1: Correlation between intellectual humility and emotional reaction to the experiment by condition
‘Shooter bias’ data:

Both experiments 2 and 3 collected data from a ‘shooter bias’ paradigm (Correll, Park, Judd, & Wittenbrink, 2002) as a pretext for the delivery of scripted feedback. Since these data were collected at a point before any differences between the experiments or across conditions occurred, and so they can be combined (n=283). The results replicate previous studies, showing a significant difference in participants’ responses to black compared to white targets. No race-based difference was observable in reaction times, although participants were faster to respond to armed targets (means [standard deviations]: unarmed white targets 0.611 [0.047], unarmed black targets 0.612 [0.053], t(282)=0.21, p=0.832; armed white targets 0.569 [0.043], armed black targets 0.567 [0.042], t(282)=0.85, p=0.397). Accuracy — that is, getting the correct responses: ‘shoot’ (when target is armed) and ‘don’t shoot’ (when target is not armed) — revealed systematic differences according to target race. Participants were less accurate for unarmed black targets (0.67 [0.15] compared to unarmed white targets (0.70 [0.15]), a significant difference (t(282)=3.55, p<0.0001). This represents a greater likelihood of shooting an unarmed target if they are black. Participants were more accurate for armed black targets (0.74 [0.13]) than for armed white targets (0.72 [0.12]). This represents greater likelihood of correctly shooting an armed target if they are black. This result provides a replication of the shooter bias effect with a UK based sample. This is particularly interesting because regulation and attitudes towards carrying and using firearms are substantially different in the UK compared to the USA where most of the existing work using this paradigm has been conducted.
Additional measure conducted at the end of experiment 2
Experiment 2 included a CV task based on two existing studies into hiring discrimination (Uhlmann and Cohen 2007; Lindner, Graser and Nosek 2014). Participants were asked to rate the suitability of a candidate for a particular job. The jobs were police chief, company representative, and factory manager. Participants received a sheet describing some of the qualities and qualifications of the candidate which also included a picture. For half the candidates these pictures were of white males and for half the candidates the pictures were of black males. The police chief was deliberately designed as an unsuitable candidate in order to test for the demand characteristic of participants deliberately choosing to rate any black candidate highly. We tried analysing the results in a number of ways including an analysis which excluded those who rate the poor candidate highly but the only evidence of bias is in favour of black candidates. These results are probably best explained by demand characteristics.

Additional measures at the end of experiment 3
Experiment 3 contained two additional measures which were not reported in the main write up. These measures took place at the end of the experiment after the measures reported in the main write up but before the experimental debrief.

The first of the additional measures was a seating distance measure. This was included to try to get a behavioural measure that goes beyond merely clicking on a computer screen. Participants were told: “These computer tests are only one aspect of people’s racial attitudes and associations. We are also interested in how they relate to actual social interactions. To investigate this we have organised for you take part in a pair problem solving task with Jamal a black British student. Don’t worry it is nothing too difficult and it will only take a few minutes”. Then shown to a room. Layout shown in figure 3 below. On entering the room participants are told. “Oh, Jamal seems to have gone somewhere. His stuff is still here. So he won’t have gone far. Take a seat and I'll go find him.” Their seating choice was recorded when the experimenter returned.

**Figure 3**- Top down view of the layout for the seating distance measure.

The results did not show anything of significance. There was no difference in means between conditions: Control 2.47 (n=79) vs Blame 2.55 (n=76) and no correlation between seating distance and IAT score: 0.07.

The second additional measure was a waiting measure. When the experimenter returned to the room where the participants were ‘waiting for Jamal’ he told participants “I’m really sorry but Jamal seems to have popped out for some reason. I’ll give him a call. How long are you able to wait?” The participants full answer including any excuses offered were recorded by the experimenter. The results of this measure did not show anything of
significance. There was no difference in mean waiting time between the two conditions. Control mean: 48.02 minutes (unlimited/no answer 30), SD: 45.44; Blame mean: 44.20 minutes (unlimited/no answer 27), SD: 37.75. This measure turned out to be very difficult to implement as many participants were happy to wait for an unlimited time or arrange to come back, or insist that the experimenter try calling him several times to find out how long he will be or ask a number of questions. Many of those who gave answers in minutes said they could only wait that long because that is when their next class starts.

References: