Racial bias in face perception is sensitive to instructions but not introspection

Eoin Travers[1 2]

Merle T. Fairhurst[1 3]

Ophelia Deroy[1 3]

[1]Centre for the Study of the Senses, School of Advanced Study, University of London

[2]Institute of Cognitive Neuroscience, University College London.

[3]Munich Center for Neuroscience, Ludwig-Maximilian University, Munich

Address for correspondence
Eoin Travers
Institute of Cognitive Neuroscience
University College London
WC1N 3AR, UK
Email: e.travers@ucl.ac.uk

Abstract

Faces with typically African features are perceived as darker than they really are. We investigated how early in processing the bias emerges, whether participants are aware of it, and whether it can be altered by explicit instructions. We presented pairs of faces sequentially, manipulated the luminance and morphological features of each, and asked participants which was lighter, and how confident they were in their responses. In Experiment 1, pre-response mouse cursor trajectories showed that morphology affected motor output just as early as luminance did. Furthermore, participants were not slower to respond or less confident when morphological cues drove them to give a response that conflicted with the actual luminance of the faces. However, Experiment 2 showed that participants could be instructed to reduce their reliance on morphology, even at early stages of processing.

Note: All stimuli used, code to run the experiments reported, raw data, and analyses scripts and their outputs can be found at https://osf.io/brssn/?view_only=f458374cf57d4d54b231f9cac252ce95, to be made public on publication.

Keywords: Perception; Metacognition; Confidence; Social cognition;

Exploring the Mechanisms of Racial Bias in Face Perception

Perception is shaped by prior expectations (Bruner, 1973; Kersten, Mamassian, & Yuille, 2004; Stocker & Simoncelli, 2007), which in some cases leads to systematic distortions and illusions (Gregory, 1968; Morgan, Hole, & Glennerster, 1990). In many cases, these biased percepts are of little consequence: they are only seen under laboratory conditions, and often only affect our perception of impoverished, abstract stimuli such as lines and patches of colours. Some, however, affect our perceptions of everyday stimuli. Levin and Banaji (2006) demonstrated one such bias in the perception of faces. They showed that faces with typically African features are perceived as darker than those with European features of the very same luminance. Similarly, manipulating a face's hairstyle to be typically African American or Hispanic affects people's judgements about the face's complexion (MacLin & Malpass, 2003). The effect of ethnic categorisation on perception of complexion is often described as a racial bias, and will be referred to here as the race-lightness effect. This bias is of particular interest as it bridges gaps between two disciplines where biases are well-studied: perception (e.g. length estimation arising from the Müller-Lyer illusion; Morgan, Hole, & Glennerster, 1990), and social cognition (e.g. those stemming from implicit associations; Greenwald & Banaji, 1995).

The mechanisms underlying the race-lightness effect remain unclear. Possible explanations include top-down influences from category-level representations (Levin & Banaji, 2006), bottom-up correlations between features encoded in the perceptual system (Firestone & Scholl, 2014, 2016), or, as recently proposed, dynamic interactive models (Freeman & Ambady, 2011a; Freeman, Johnson, Adams, & Ambady, 2012). We aimed to clarify three key characteristics of the process driving this bias: its temporal unfolding, its accessibility to metacognitive awareness, and its sensitivity to high-level intentions.

First, how early does the bias emerge? It may be that initial perception of luminance is distorted by morphological cues as soon as luminance is processed at all. Alternatively, the luminance might be initially processed independently and accurately, but is subsequently distorted by morphological cues after further processing. Most perceptual biases seem to show the first profile. In the Müller-Lyer illusion, for instance, perceivers do not initially see both lines as being of the same length, and only come to believe that one line is longer than the other as they process the other geometric features. There is also good evidence that most social biases, including inappropriate stereotype usage (Macrae & Bodenhausen, 2000) and implicit associations (Greenwald, Mcghee, & Schwartz, 1998), occur quickly. For instance, these biases are more pronounced when participants perform tasks under time pressure, and they affect response times on

tasks for which they are not relevant (Greenwald et al., 1998). On the other hand, the race-lightness effect is a somewhat unusual perceptual bias, in that it involves the integration of two visual dimensions that are at least partly processed independently – shape/edge detection, and luminance/colour (Zeki, 1993). Some have also proposed that the bias is mediated by category-level representations (Levin & Banaji, 2006). If this is true, we would expect the bias to only emerge once an initial categorisation decision - "White" or "Black" - has been made.

Second, to what degree are participants aware of this bias? Much recent work (Deroy, Spence, & Noppeney, 2016; Fleming & Frith, 2014; Shea et al., 2014) has explored the role of metacognitive awareness in perceptual decision making: participants' awareness of the processes underlying their own judgements, and their awareness of the accuracy of these judgements. Data from post-decision confidence ratings are commonly used to infer awareness of this sort, and congruent with evidence obtained from response times, bets, or resampling decisions. The relationship between metacognitive awareness and confidence is discussed by Rausch, Müller, & Zehetleitner (2015). Explicit confidence ratings have been used, for instance, to show that unconsciously processed information can drive decision making without affecting subsequent confidence ratings (Vlassova, Donkin, & Pearson, 2014). Using confidence as a measure of metacognitive awareness in this way allows us to avoid some of the demand characteristics associated with directly asking for explicit self-reports (Rausch, Müller, & Zehetleitner, 2015). While metacognitive awareness is typically studied in the perception of a single property, it can also be studied for integrated percepts, when two cues are combined (Deroy et al., 2016).

Particularly relevant here is the evidence that participants can be metacognitively aware of the conflict between cues which explains certain multisensory illusions. White et al. (2014) presented participants with variants of the well-known McGurk effect (McGurk & MacDonald, 1976). This effect shows that conflicting visual and auditory verbal stimuli may lead participants to experience a new illusory integrated phoneme, for instance a /da/ when hearing a phoneme /ba/ and seeing an actor's lips articulating a /ga/. Notably, White et al. showed that participants were less confident in their judgments when the /da/ percept was produced by combining conflicting cues (auditory /ba/ and visual /ga/) than when reporting the same percept produced by consistent cues (auditory and visual /da/). This indicates that participants had some access to the relation between the underlying cues, and not just their final composite. Here, we test whether the same holds for the race-lightness effect. If morphological cues are automatically integrated into the processing of luminance at an early stage, participants may not be able to metacognitively assess their error. Alternatively, although participants' responses are based on a combination of these two cues, they may maintain residual independent representations of each, which can contribute independently to participants' post-decision confidence ratings.

A more drastic possibility is that the race lightness effect does not involve perceptual integration at all. Instead, participants may have access to both cues, but simply err when producing their responses. This appears to be the case for at least one social bias. Payne, Shimizu, and Jacoby (2005) demonstrated that participants who misperceive tools as guns when held by African Americans are able to detect and correct their errors when given sufficient time. This indicates that this particular bias stems from a failure to control rapid, automatically produced responses, rather than a persistent visual illusion. The same phenomenon occurs on many other perceptual and cognitive tasks requiring participants to ignore a conflicting cue (Eriksen, 1995; Stroop, 1935). If this were to be the case for the race lightness effect, we would expect to see a dramatic dissociation between the cue weightings that drive participants' responses and those that drive their post-decision confidence ratings.

A third and final question is whether this bias is affected by higher level goals and intentions. Whether low-level perceptual processes are affected by higher-level states is a divisive issue. Modularist views consider that lower level processes take place in encapsulated neural modules (Fodor, 1983), which only pre-process perceptual information before passing it on to higher level cognitive systems, where beliefs are formed (Firestone & Scholl, 2016). The alternative view is that information is continuously exchanged between so-called cognitive and perceptual processes (Freeman & Ambady, 2011a; Friston, 2010; Lupyan, 2015). As a result, perceptual information continuously affects cognitive states, and vice versa. There is evidence that prior beliefs and expectations do affect visual processing (Lupyan, 2015), although it remains uncertain at what stage in processing these effects emerge. Indeed, this is one possible explanation for the race-lightness effect – faces with African features appear darker because the perceiver categorises the face as being African, and the expectation that African individuals have darker skin propagates back to the visual system (Levin & Banaji, 2006). If the bias is indeed driven by such top-down cognitive effects, it likely can be reduced by top-down intentional control. The effect of instructions on perceptual and cognitive processes have been discussed elsewhere (Brass et al, 2017; Hertz, Blakemore & Frith, 2020). Interestingly, performance on the IAT does not improve when participants are asked to avoid or suppress racial biases (Lai, Hoffman, & Nosek, 2013; Lai et al., 2014). It is worth considering, however, that the biases measured by the IAT act on rapid cognitive evaluations, whereas the race-lightness effect acts on lower-level percepts. We are not aware of any previous work testing whether high-level instructions affect participants' bias towards misperceiving items held by African Americans as weapons.

If there is an intentional modulation of the effect, it is also important to consider *when* higher level intentions alter this effect. If they simply make participants retrospectively aware that their judgements are biased, they should only alter the pattern of post-decision confidence ratings. If

higher-level control makes participants more cautious in their responses, participants should respond more slowly when they are made aware of their bias, as they attempt to override it. The third option is that high-level intentions can affect the way faces are perceived in the first place. In this case, we would expect to find a diminished effect of morphological cues even in early mouse cursor movements, as participants decide between response options.

To answer these questions, in Experiment 1 we presented participants with a task based on those used by Levin and Banaji (2006; see Figure 1). Participants saw two computer-generated faces in sequence, for 750 msec each, and were asked to indicate if the second face (the target) was lighter or darker than the first (the reference; Figure 1b). We varied both the morphological features of the faces (typically African, ambiguous, or typically European) and their objective luminance (darker, neutral, or brighter), in order to estimate the contribution of both morphology and luminance to participants' responses, as well as their mouse cursor movements towards each response option over time, their response times, and their subsequent confidence ratings about their decisions. This two-interval forced choice-design differed from the paradigms used by Levin and Banaji (2006) who primarily asked participants to match the luminance of faces with different morphological features (their Experiments 1-3). We chose this design as confidence ratings are difficult to formulate with a continuous estimation task like that used by Levin and Banaji (2006), but easily applied to binary forced-choice tasks (Fleming & Lau, 2014).

Our hypotheses were as follows. First, we expected our analysis of participants' responses to replicate Levin and Banaji's results, using a novel paradigm: responses should be driven not only by the actual luminance of the faces, but also by the morphological cues. Second, the mouse cursor trajectories provide a window into the early temporal profile of the bias (Figure 1c). Specifically, they make it possible to infer at what point in time participants' movements were first influenced by luminance and by morphological cues, and how strong these influences were, over time (Freeman & Ambady, 2011b; Travers, Rolison, & Feeney, 2016). Our focus here is whether the earliest effect of morphological information on participants' movements can be seen at the same time as the earliest effect of luminance, indicating an early, and likely perceptual source to the bias, or some time afterwards, indicating a later, perhaps cognitive mechanism. Mouse tracking has been extensively used in previous studies of the temporal dynamics of face perception (Freeman & Ambady, 2009, 2011). In the standard task, participants click a button located in the bottom centre of the screen to reveal a face stimulus and are required to quickly click on one of two possible response options (e.g. "Male" or "Female"), located in the upper corners of the screen. Rather than moving directly towards one or other option, participants typically trace a curved path towards their response, with the attraction towards the non-selected option increasing as participants experience greater response

conflict. This curvature, and the timing of tentative movements towards either option, provide a window into the processes that eventually produce an overt response.

Third, we use confidence ratings to assess participants' access to the individual perceptual features of the stimuli. Our approach here is as follows. We assume that participants' perception of facial luminance is a weighted combination of the objective luminance and the available morphological cues. To produce a response, participants subtract the perceived luminance of the target face from that of the reference, and respond "Lighter" if this difference exceeds a criterion, with Gaussian noise impairing this decision. The cue weightings and threshold used by each participant can be estimated using probit regression. Given these weightings, we can infer how close to criterion the perceived luminance of each pair of faces is, and thus predict expected confidence if participants have access to only the difference in *perceived* luminance. Our question is therefore whether participants' reported confidence is driven only by how close the stimuli were to criterion on each trial, or if it is also influenced by the degree to which objective luminance is congruent with the response given.

As an example, consider a participant who gives equal weighting to both objective luminance and morphology when perceiving luminance, $B_{Luminance} = B_{Morphology} = 1$, where B_{Cue} is the weight given to each cue. On a trial where the target is two levels lighter in luminance, but has features that appear one level more African, the perceived luminance is $(2\times B_{Luminance} - 1\times B_{Morphology}) = (2\times1 - 1\times1) = 1$ standard deviation above criterion. Likewise, on a trial where the target is one level darker in luminance and features that appear two levels more European, perceived luminance is again $(-1\times1 + 2\times1) = 1$ standard deviation above criterion. Assuming no bias, the probability of responding 'Lighter' in both cases is therefore $\Phi(1)$, where Φ is the Gaussian cumulative distribution function. Importantly, if the participant lacks metacognitive awareness to the original cues, distance from criterion should be the only factor that influences their confidence, and so confidence should be matched for both trials. On the other hand, if they maintain residual independent access to the cues, they should be more confident in the first case, where the true luminance provides greater supports for their response (Figure 1d).

Finally, in Experiment 2, we informed participants halfway through the task that their responses were biased and asked them to try to avoid this. This allowed us to repeat the analyses described above, both before and after the instructions, and infer what aspects of participants' behaviour – their responses, confidence ratings, response times, and early cursor movements – were changed, if any.

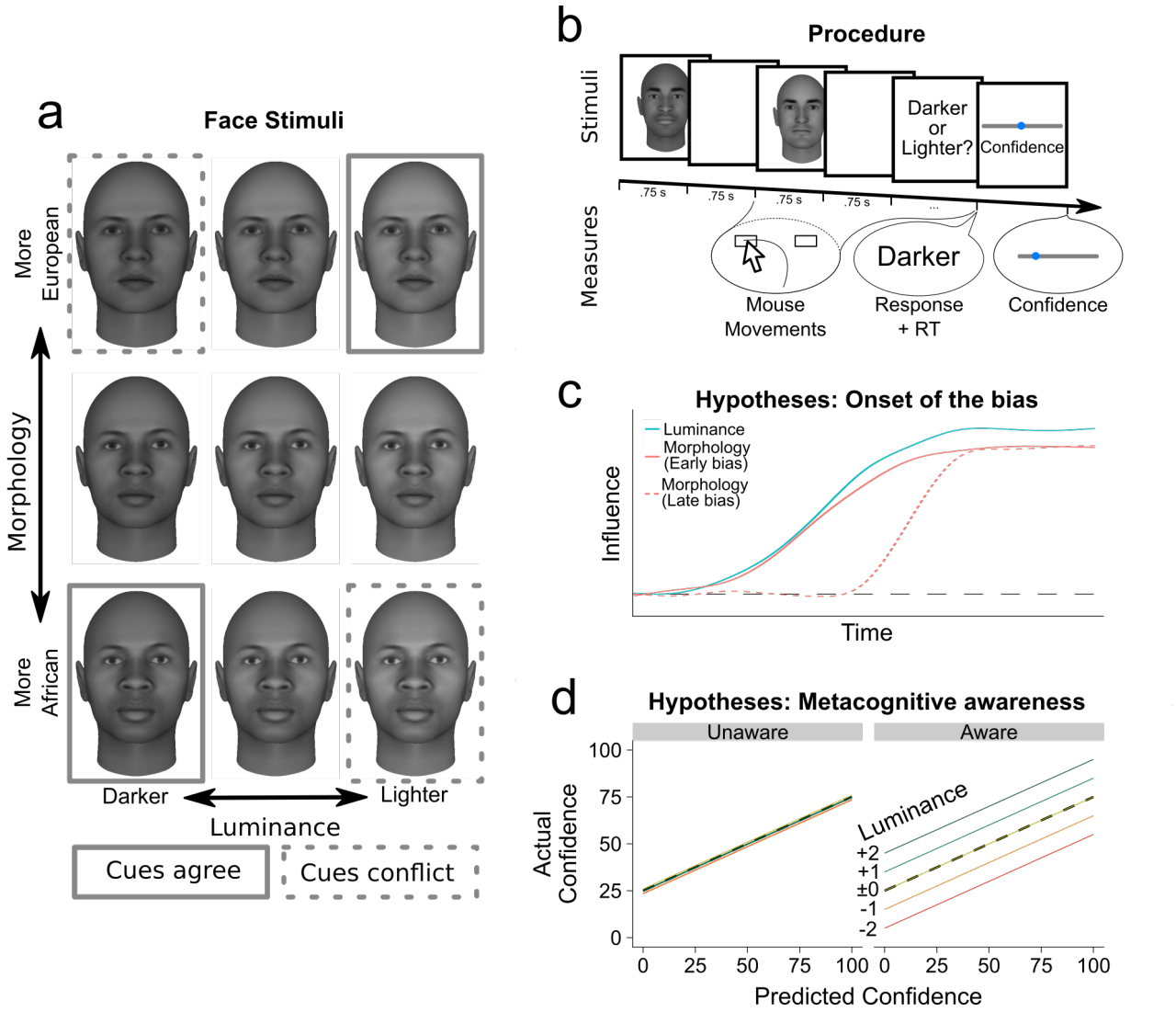

Figure 1. a) Nine versions of a face generated for the experiment, varying in luminance (horizontal axis) and morphology (vertical). For some pairs of faces, luminance and morphological cues support the same response (solid boxes), while for others they conflict (dashed boxes). b) Experimental task. Participants were asked to indicate if the second face was darker or lighter than the first. We recorded participants' responses, their mouse cursor movements as they selected between the two options, and their subsequent confidence ratings. c) The influence of morphology on participants' cursor movements may either begin early and concurrent with the the influence of luminance (solid red curve) or later (dashed red curve). d) Participants' confidence ratings may use only the luminance cues that drove their responses (left), or show additional awareness of whether or not luminance cues were consistent (positive values) or inconsistent (negative values) with the response given (right). Stimuli were generated using FaceGen Modeller (version 3.14; https://facegen.com/modeller.htm)

Experiment 1

Methods

All experimental protocols were approved by the University of London School of Advanced Study ethics committee, and all tests were carried out in accordance with the approved guidelines. All participants completed online informed consent forms after reading instructions outlining the experiments.

Participants

Participants (N = 77, 40 female) were recruited online using Mechanical Turk. In the most relevant previous study, using similar stimuli to those used here, Levin and Banaji (2006) asked participants to adjust a target to match the luminance of a face shown onscreen, and reported effect sizes between d = 1.15 and d = 1.65. A sample size of 80 would have 80% power to detect a much smaller effect, of d = 0.3. Additionally, in our task participants were asked to compare two faces, one of which was held in memory, and so could not view both stimuli at the same time. We would therefore expect the bias here to be more pronounced. Our use of mixed models, with continuous predictors (see below) also serves to increase our statistical power. Therefore, although we primarily report Bayesian parameter estimates throughout this manuscript, rather than null hypothesis significance tests, this power calculation indicates that a sample of 80 participants would allow us to measure the effect of morphological cues on luminance judgements with excellent sensitivity. We therefore aimed to test 80 participants, but did not acquire data from 3 who claimed but did not complete the task on Mechanical Turk.

Sixty-six participants stated they lived in the USA, eight in India, and one each in Guyana and Israel. We also asked participants if they identified as members of a particular ethnic group. Of those in the USA, forty-seven identified as White or equivalent, and the remainder as Black (4), Asian (4), mixed race or any other group (4), or did not say (7). The average age was 37 (SD = 9). The mean age was 37 years (SD = 9).

Stimuli

Using FaceGen (Singular Inversions, 2016), we generated eight unique male faces, and produced three versions of each, with facial features (morphology) either a) typically African, b) ambiguous, or c) typically European. Following Levin and Banaji (2006), to match the luminance of these images, we converted them to greyscale, algorithmically adjusted the whites of the eyes so that

pixels more than two SDs brighter than the rest of the face were systematically darkened, and then adjusted the entire set of images to their mean luminance (95/255, 8-bit encoding). We then adjusted the luminance of these standardised greyscale images to create three luminance levels: a) 90% brightness (84/255), b) 100% brightness, and c) 110% brightness (103/255). These brightness levels were chosen so that a difference of one level (90% vs 100%, or 100% vs 110%) could be discriminated, but with less than perfect accuracy. Thus, we used 72 images in total: 8 unique faces, at 3 levels of morphology, and 3 levels of luminance. The final stimuli used in this experiment, as well as the source images and preprocessing scripts used to generate them, are available in the Open Science Framework repository accompanying this article (https://osf.io/brssn/?view_only=f458374cf57d4d54b231f9cac252ce95).

We generated 144 pairs of images, with two different faces in each pair. On each pair, we coded the difference in morphology, and the difference in luminance, between reference and the target. For morphology, typically African faces were coded [-1], neutral faces as [0], and typically European faces as [+1]. As an example, this means that a trial with a typically African reference and a typically European target was coded as a difference of [+2], whereas a trial with a neutral reference and a typically African target was coded as [-1]. Similarly, for luminance, darkened faces were coded as [-1], unaltered faces as [0], and lightened faces as [+1]. We factorially crossed the 5 possible levels of luminance difference [-2 to +2] with the 5 possible levels of morphology difference, and included 6 trials for each combination, excluding pairs where both luminance and morphology were matched for both images. We also balanced the number of trials in which the target and the reference were morphologically African, neutral, or European, and the number where the target and the reference were darkened, unaltered, or lightened. The order of presentation was randomised for each participant.

Procedure

On each trial, participants were shown two faces, in sequence, and asked to state whether the second face (the target) was darker or lighter than the first (the reference). After responding, they were asked to indicate how confident they were that their response was correct (Figure 1B). The faces were presented in the centre of the screen on a white background and subtended 20% of the height of the window. Participants saw a fixation cross, and each face, for 750 msec each, with a 750 msec inter-stimulus interval between them. 750 msec after the offset of the target face, buttons marked “Darker” and “Lighter” appeared in the top corners of the screen, and a prompt reading “Darker or Lighter” was shown centrally. The left/right location of the two buttons was randomised across participants. Once either button was clicked, participants reported their confidence by dragging a continuous slider between ends marked “Not certain at all” (left) and “Totally certain”

(right), before clicking on the "Next" button, in the bottom centre of the window to initiate the next trial. Each block of trials was initiated by clicking a "Go" button, in the same location.

Participants were told at the start of the experiment that their task was to decide if the second face was darker or lighter than the first, and the confidence scale was explained. They were also told that if they used the confidence scale properly, they should be more confident when they are right than when they are wrong. Participants then completed 20 practice trials, using versions of the stimuli where the pixels of the face had been scrambled to produce grey silhouettes with the same overall luminance. After the practice trials participants were told how many correct responses they gave, and their average confidence for correct and incorrect responses. They then proceeded to the 144 experimental trials.

Modelling Approach

To analyse the mouse cursor data, we normalised the cursor positions to a standard x-y co-ordinate system. The initial cursor position was normalised to [0,0], the width of the display to 2 units, and the height to 1.5 units, preserving the true aspect ratio used in the experiment. We then re-sampled the horizontal cursor co-ordinates in 20 msec intervals. As response times differed between trials, we padded the data to a common length of 3 seconds by adding extra samples of the final cursor position. To estimate the influence of luminance and morphology cues over time, we fit a linear mixed model to the data from each 20 msec window, and extracted the regression weights and p-values from each. If the influence of morphological cues arises early in processing, morphology should affect participants' cursor movements from the same point at which luminance cues do. If it arises only after subsequent processing, morphology should begin to affect participants' cursor movements at a somewhat later time than luminance (see Figure 1c).

To estimate the influence of each cue on participants' responses, we fit a probit mixed model, using the lme4 package for R. The model included an intercept parameter capturing the bias toward responding 'lighter', and main effects capturing the influence of the difference in morphology and the difference in luminance between the two images (using the coding scheme above, from -2 to + 2). All predictors were included both as fixed effects, indicating the average influence of each cue across all participants, and random effects, indicating the influence for each individual participant. Note that this approach differs from typical previous work using mouse-tracking to explore social perception (e.g. Freeman & Ambady, 2009). In these studies, participants make judgements with near perfect accuracy, such as indicating whether a face is male or female. The analyses then focus on the different between cases where a secondary cue conflicts with one,

and test whether participants show greater signs of uncertainty before producing a correct response on these conflict trials. Here, on the other hand, our secondary cue (morphology) exerts a considerable influence on participants' final responses. We therefore use mouse-tracking across all trials, regardless of the final response, to explore when each cue begins to influence motor output.

For our analysis of participants' confidence ratings, we first used the probit model above to predict the value of the decision variable δ on each trial, corresponding to each participants' subjective difference in luminance between the two faces shown. This was done using R's 'predict' function. To transform this metric into predicted confidence ratings, we assume that participants follow the optimal Bayesian strategy, where the probability that a response is correct, e.g. the probability that the target is lighter if $\delta > 0$, or darker if $\delta < 0$, is simply $\Phi(\delta)$, where Φ is the Gaussian cumulative distribution function, if the participant responded 'lighter', and $\Phi(-\delta)$ otherwise (see Hangya, Sanders, & Kepecs, 2016). Finally, we fit a linear mixed model to participants' actual confidence ratings. As predictors, we include the predicted confidence ratings, and the actual luminance evidence on each trial, ranging from supporting the response given by two levels (+2) to conflicting by two levels (-2). If participants maintain access to the true luminance of the stimuli, the luminance evidence should predict confidence ratings over and above the level of confidence expected from responses alone. On the other hand, if participants do not have access to this information, it should not influence their confidence ratings (see Figure 1d). We repeat this analysis for participants' log-transformed response times, following the same principles. We also report models fit to participants' raw confidence ratings and log-transformed response times, with the intercept term, luminance and morphology cues in support of the response given, and progress through the experiment (first trial = -0.5, last trial=+0.5) included as predictors. All predictors were included as both fixed and random effects.

Results

Data Screening and Coding

All participants' responses were at least somewhat influenced by either the luminance or the morphological cues, and so no data were excluded for this reason. Four of the 10,944 trials were lost due to connectivity issues, and we excluded 176 trials with response times (RT, measured from the onset of the second face) greater than 5,090 msec, ten times the median, 250 further trials where RT was more than three SDs above that participant's average, and data from one participant who took an average of 1,963 msec to respond after these trials had been removed, 4.6 standard deviations above the average across participants. The remaining 76 participants had an average RT of 654 msec (SD = 238).

Responses

Replicating Levin and Banaji's main finding (Figure 3a), we found that participants responses were driven by both luminance cues, B = 0.83, CI = [0.74, 0.92], z = 17.76, p < .001, and morphology, B = 0.61, CI = [0.51, 0.72], z = 11.51, p < .001. Both parameters varied considerably across participants. The estimated SD of the effect of luminance across participants was 0.15, with a range across participants between 0.04 and 1.63. The SD for the effect of morphology was 0.20, with a range of range between -0.26 and 1.61. There was also a significant negative intercept term, B = -0.41, SD across participants = 0.43, CI = [-0.52, -0.30], z = 7.62, p < .001, indicating an overall bias towards responding that the target was darker than the reference.

Early Temporal Profile

Next, we used the mouse cursor data to test when each of these cues first began to influence participants' actions, by inspecting the horizontal position of the cursor over time. This tracks how close the cursor was to the 'Darker' and 'Lighter' response options. Our analysis begins at the onset of the second face (0 msec) and continues through its offset (750 msec) and the onset of the response prompt (1,500 msec). The response options were shown on screen throughout this time. The average response time from the onset of the second face was 2,239 msec (SD = 304 msec; Figure 2b), and participants on average initiated their cursor movements 1,126 msec (SD = 319 msec; Figure 2a) after the onset of the second face, or 374 msec before the onset of the response prompt.

First, we analysed the data collapsing across all responses. This revealed the contribution of each cue to behaviour over time. It should be noted that the data from the end of the time window

here correspond to participants' ultimate responses, and the corresponding regression weights match those reported in the analysis of these responses, above. The data from earlier times reflect participants' tentative movements towards the response options. As a consequence, the results of this analysis at – for instance – 1.5 seconds should correspond to the pattern of responses we would expect to see if participants were forced to respond at this time. Figure 2c shows the average cursor position from subsets of trial where the target was either much darker or much lighter than the reference, and the morphological cues either strongly agreed or strongly conflicted. Figure 2d shows the influence of both kinds of cue on cursor position, estimated at 20 msec intervals. Luminance significantly influenced participants' cursor movements from 460 msec after the onset of the target face (i.e. responses to the lighter and darker stimuli diverged). Importantly, morphology also significantly influenced cursor movements from 460 mseconwards (responses to conflict and no conflict stimuli diverged in Figure 2c). Furthermore, the effect of luminance and the effect of morphology were tightly linked over time, and only began to differ significantly from each other much later, at 1,320 msec (when the error regions in Figure 2d no longer overlap). Therefore, the influence of the morphological cues arose almost concurrently with the influence of the luminance cues, evidence that that the bias emerges relatively early in visual processing.

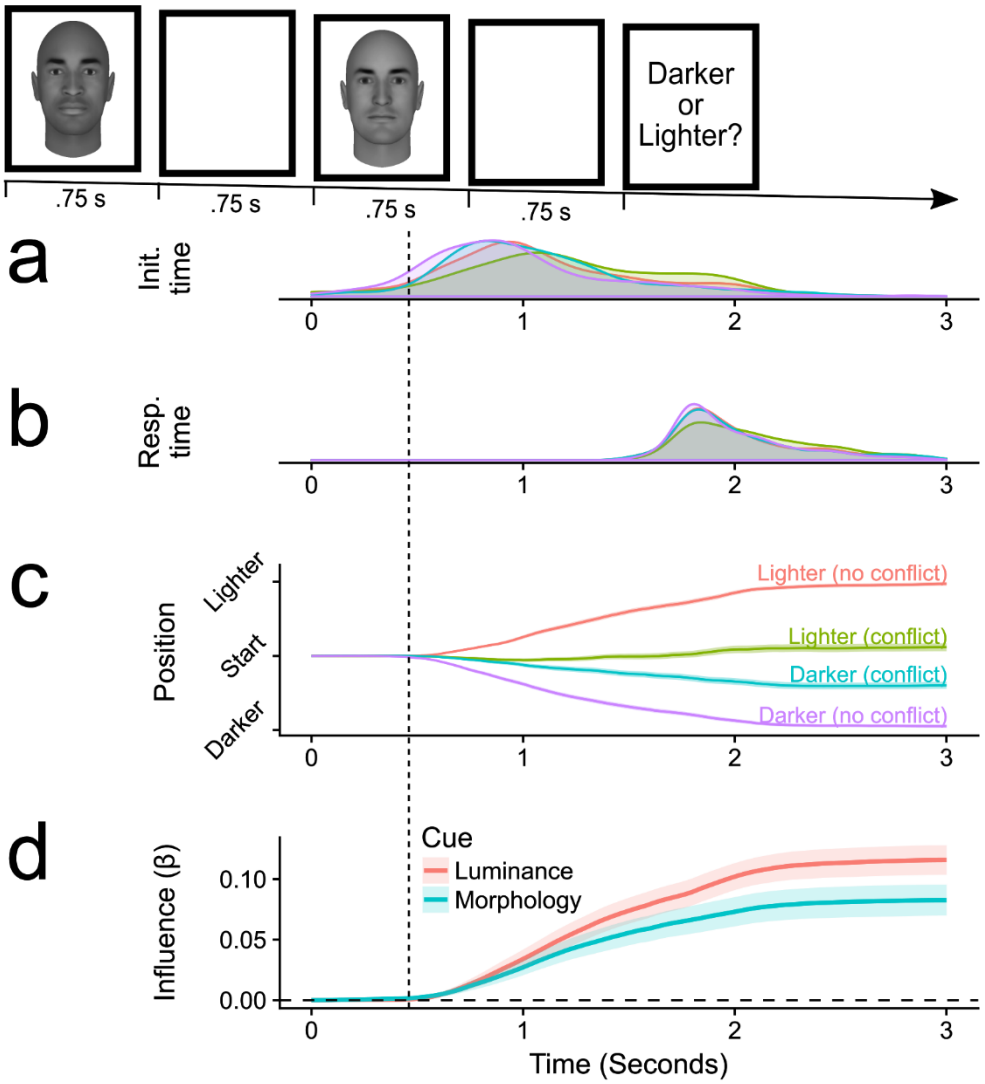

Figure 2. Mouse cursor trajectories from Experiment 1 show the early temporal profile of the race-lightness effect. Time line shows stimulus presentation times, with the onset of the target face coded as time 0. a) Movement initiation times. b) Response times. c) Mean x-axis position of the mouse cursor over time, for the subset of trial trials where the target face was two levels darker or lighter than the reference, and the morphological cues either agreed or conflicted with the luminance. d) Regression weights over time, indicating the influence of luminance and morphological cues on cursor position. Both cues significantly influenced cursor position from 460 msec after target onset onwards (vertical dashed line). Shaded regions in c and d show standard error. Face stimuli were generated using FaceGen Modeller (version 3.14; https://facegen.com/modeller.htm).

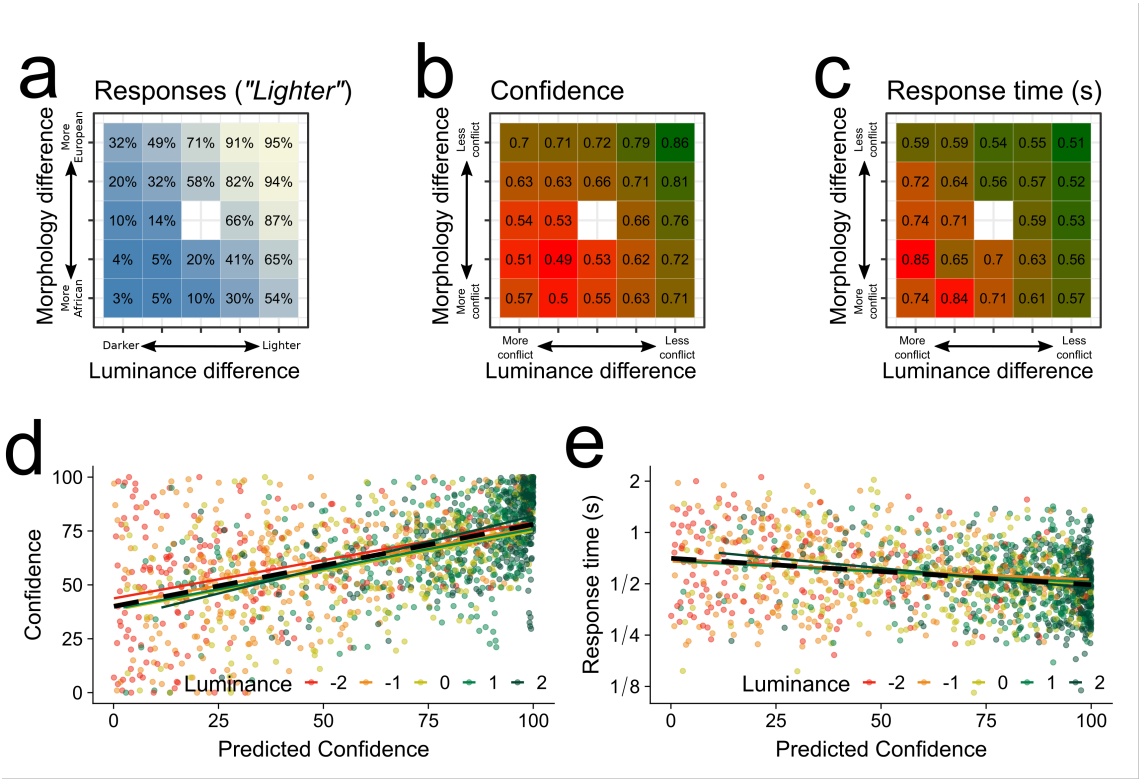

Figure 3. Responses, confidence ratings, and response times from Experiment 1. a) The probability of participants stating the target was lighter than the reference was predicted by the differences in luminance and in morphology between the images. b) Confidence ratings, and c) response times were also predicted by the degree to which the differences in luminance and morphology supported the response given. d) Reported confidence ratings were linearly related to predicted confidence, derived by estimating the distance of the perceived difference in luminance from criterion on each trial. After taking this index into account, confidence was not additionally predicted by the objective luminance of the stimuli presented on each trial. e) The same result holds for response times.

Metacognitive Awareness

Response times and confidence ratings were negatively correlated within each subject, mean r = -0.37, SD = 0.19, d = 1.93, CI = [1.55; 2.32], t(74) = 16.759, p < .001, t-test across participants, and participants with a faster average response time also had a higher average confidence, r(73) = -0.42, p < .001. Therefore, it seems these measures provide similar indices of conflict or uncertainty during judgement. In what follows, we report analyses for both measures.

Like their responses, participants' confidence ratings were influenced by both luminance cues, B = 7.30, CI = [6.06, 8.54], t(71.2) = 11.579, p < .001, SD across participants = 5.23 , and morphology, B = 5.28, CI = [4.10, 6.46], t(70.3) = 7.558, p < .001, SD across participants = 5.02. Overall, confidence did not change over the course of the experiment, B = 0.85, CI = [-1.42, 2.55], t(74.0) = 0.561, p = .577, but with notable individual differences, SD across participants = 6.63, individual participants' estimates from -13 (e.g. confidence 13 points lower at the end of the experiment than the start) to + 17. Response times (log transformed for analysis) were also influenced by both luminance cues, B = -0.10, CI = [-0.08, -0.12], t(64.0) = 9.633, p < .001, SD across participants = 0.07, and morphology, B = -0.07, CI = [-0.05, -0.08], t(68.0) = 7.714, p < .001, SD across participants = 0.06, RTs also become significantly faster as the experiment progressed, B = 0.15, CI = [-0.09, -0.21], t(73.8) = 4.808, p < .001, SD across participants = 0.21.

Figure 3d shows participants' reported confidence ratings as a function of their predicted confidence, separately for each level of objective luminance, from -2 (luminance conflicts with the response given by two levels) to +2 (luminance supports response given by two levels). As expected, our predicted confidence measure was significantly related to participants' actual confidence ratings, B = 47.1, CI = [41.1, 53.0], t(75.1) = 15.507, p < .001, SD across participants = 24.5. Objective luminance had no effect on participants' confidence ratings over and above what was predicted from participants' responses alone, B = 0.3, CI = [-0.4, 1.0], t(31.7) = 0.932, p = .359, SD across participants = 2.9. The Bayes Factor (BF) for the effect of luminance was calculated using the Bayesian Information Criteria for models with and without this term (that is, $BF_{10} = \exp(.5 \times (BIC_0 - BIC_1))$ and revealed a BF of 66.2 in favour of the null hypothesis.

Figure 3e shows participants' response times (on a log scale) as a function of the same factors. As expected, RTs were negatively related to predicted confidence, B = -0.73 CI = [-0.82, -0.65], t(66.5) = -16.930, p < .0001, SD across subjects = 0.28. Objective luminance did not additionally predict participants' response times, B = 0.00, CI = [-0.01, 0.01], t(45.9) = 0.510, p = .610, SD across subjects = 0.02, BF_{01} (in favour of the null) = 89.5.

Discussion

In this experiment, we replicated the core finding of Levin and Banaji (2006), showing that judgements about the perceived luminance of faces are also influenced by morphological features. Our results also reveal a number of the characteristics of this bias. Our analyses of early mouse cursor trajectories are consistent with the idea that the bias arises at an early point in processing, and

in fact arises as early as the effect of luminance itself. Our analysis of participants' confidence ratings and response times indicate that participants do not have access to the true luminance of the face stimuli, unbiased by the morphological cues, or at least that they do not detect when the race-lightness effect causes them to make incorrect judgements.

Experiment 2

Having shown that this bias arises quickly, and that participants did not show evidence of having access to the underlying physical luminance of the stimuli, we next asked whether the bias could be affected by higher level knowledge and goals. In a second experiment, participants were informed halfway through that their responses were biased and asked to try to avoid this. As noted above, if instructions do reduce the bias, there are two likely ways in which this could occur. It could be that participants' actual responses are unchanged, but they show greater sensitivity to true stimulus luminance, as indexed by their confidence ratings and response times (see Deroy et al., 2016; White et al, 2014). Alternatively, they may come to rely on morphology less even in their initial responses, and so its contribution would be reduced across all measures. Additionally, if participants' responses are found to be less biased, this could be because they were less biased throughout the decision process, or, alternatively, because the bias arose as before, but was partly overridden later in processing. Mouse cursor trajectories let us differentiate between these two last possibilities. In the former case, the influence of morphology, reflected in the cursor trajectories, should be reduced as soon as it is seen. In the latter, we would expect the influence of morphology to be unchanged early on, but reduced later, as participants inhibit this cue.

We produced two intervention instructions for this experiment, along with a control condition. In the "racial" instructions, we highlighted the role of racial stereotypes in the bias. In the "featural" instructions, we simply told that the bias was driven by facial features. Our main question here is whether any instructions can alter the effect of this bias. As such, we did not make predictions about the possible differences between the two interventions, and so report the contrast between the two conditions as an exploratory analysis.

Methods

Participants

Ninety new participants were tested using Mechanical Turk (33 female, one did not report gender). In the absence of prior work which could have been used to calculate statistical power for this

manipulation, we sought to test enough participants so that the results of Experiment 1 could be replicated within each of our three conditions alone. To do so, we drew 100 random samples, of 15 participants each, from Experiment 1, and tested if morphological cues significantly affected participants' responses on each sample. We found significant effects of morphology in 100/100 cases ($p < .001$). Therefore, a sample of 15 participants should have close to 100% power to replicate the effect of morphology on participants' responses. Erring on the side of greater sensitivity, we attempted to assign 30 participants to each experimental condition here, so that we could be confident that we could sensitively measure changes in the influence of morphology. Eighty-one participants stated they were American, of whom 54 stated gave their ethnic group as White or equivalent, 7 as Black, 7 as Asian, and 4 as Latino, while 10 did not specify. The remaining participants were from India (3), Britain, Russia, Ukraine, or did not specify (one each). The mean age was 35 years (SD = 10 years).

Stimuli & Procedure

The task was identical to that used in Experiment 1, except that after 72 of the 144 experimental trials participants were presented with one of three instructions. Participants in the featural intervention condition (N = 28) were told "You showed evidence of being prejudiced on this task, so that your judgement about each face's brightness was swayed by facial features. Please try to do better on the next block of trials." Participants in the racial intervention condition (N = 29) were told "You showed evidence of being prejudiced on this task, so that your judgement about each face's brightness was affected by racial stereotypes. Please try to do better on the next block of trials." Participants in the control condition (N = 33; unequal N was due to some participants failing to complete the experiment) were told "You're halfway through the task. There are 72 trials to go. Keep it up." Participants were told to read the instructions carefully, as they would be asked a question about them that may determine whether they receive full payment for the task. This question was presented as soon as participants dismissed the instructions. In the intervention conditions, participants were asked to type into a text box what it was they were told "was affecting your judgements about each face's brightness", while in the control condition participants were asked how many trials remained.

Modelling Approach

To estimate the effect of our interventions on participants' responses, we used an augmented version of mixed models reported above. We did not find any empirical differences between our featural and racial intervention conditions. Therefore, we report analyses collapsing across these conditions. Analyses of the three conditions separately can be found in supplementary materials.

For each condition (control and intervention in the analysis collapsing across intervention types, or control, featural intervention, and racial intervention otherwise), and stage (pre- and post-instructions) we estimated separate intercepts, and regression weights for the influence of luminance and the influence of morphology. We also included trial number as a predictor, as before. Intercepts and regression weights in each stage were allowed to vary between participants as random effects. Condition was manipulated between participants, and so no parameters involving this factor could vary between participants. The model used a logistic link function.

The key quantity estimated in this model is the relative weighting given to morphological cues in producing a response, found by dividing the regression weight for morphology by that for luminance, which we denote as β. Our question is therefore whether the change seen in β following the intervention instructions is greater than that seen following the control instructions. Due to the complexity of this model, it was not possible to fit using maximum likelihood estimation as implemented in the lme4 package. We instead used Bayesian estimation, implemented in the Stan language (Carpenter et al., 2016) and the brms package for R (Bürkner, 2017). We used minimally-informative priors for both the fixed effects [B ~ Normal(0, 10)] and the random effect variance [σ ~ Half-Cauchy(0, 4)]. Finally, we conducted two analyses, one coding the racial and featural interventions as separate conditions, and one collapsing across interventions. Analyses treating each intervention separately are reported in Supplementary Materials. The full regression equations can be found in the Supplemental Materials, and R code used to fit the model is included in the OSF repository accompanying this article.

We followed a similar approach in our analysis of cursor trajectories. We again augmented the model used in Experiment 1 so that the terms for the intercept, influences of luminance, and influence of morphology were fitted separately for before and after the intervention in each condition (control/intervention in our pooled analysis, and control/racial intervention/featural intervention in our unpooled analysis). It is difficult to estimate the ratio of the influence of morphology to luminance using classical methods, and computationally prohibitive to fit a Bayesian model to each time step, so we instead focus on whether the change in influence of morphology following the interventions is greater than the change following the control instructions:

$$\delta_{Intervention} = (\beta^{After}_{Intervention} - \beta^{Before}_{Intervention}) - (\beta^{After}_{Control} - \beta^{Before}_{Control});$$

where B here is the regression weight for the influence of morphology.

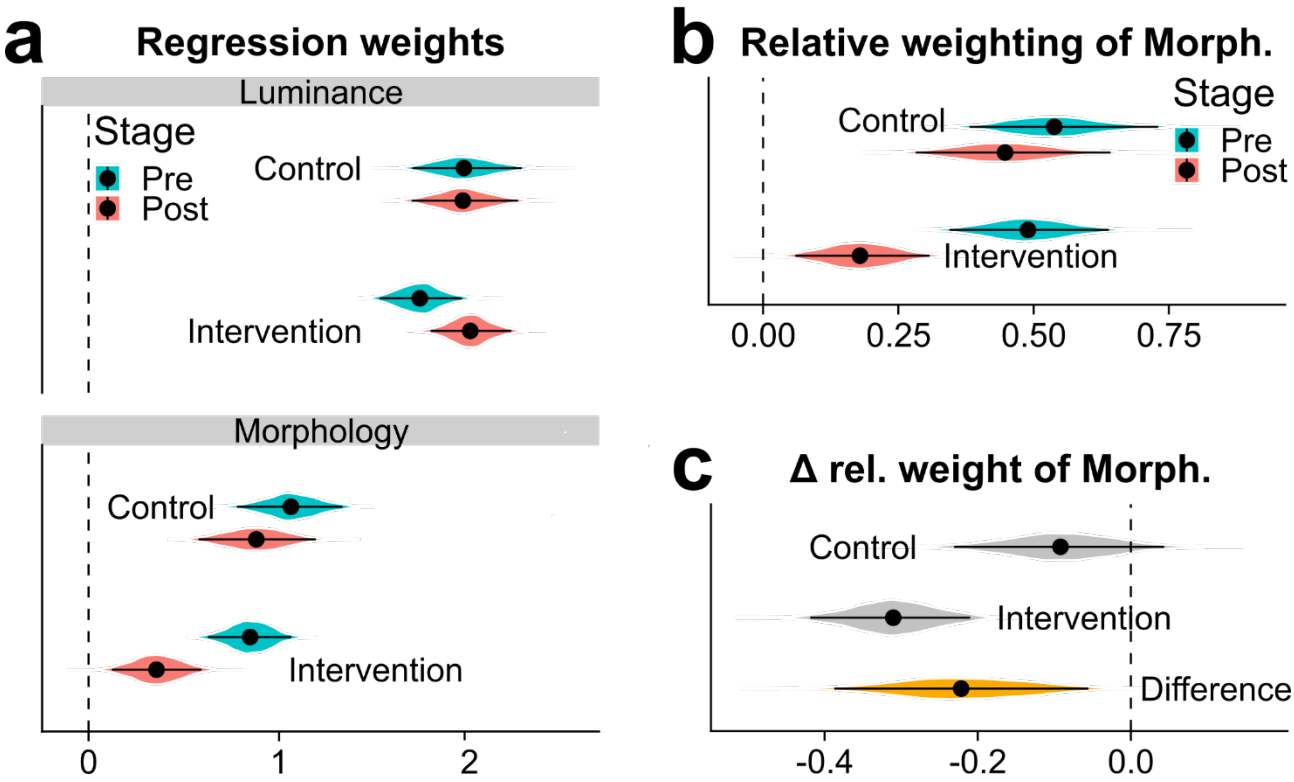

Figure 4. Posterior parameters estimates for logistic regression model fit to participants' responses in Experiment 2. Points show posterior means, error lines show 95% credible intervals, and shaded areas show posterior densities. Due to the absence of any empirical differences between them, we collapse across the two intervention conditions here. **a)** Regression weights for the influence of luminance (top) and morphology (bottom) on participants' responses, in each condition, pre- and post-instructions. The intervention instructions reduced the influence of morphology, and slightly increased the influence of luminance. **b)** Relative weighting of morphology, found by divided the influence of morphology by that of luminance, was reduced following the intervention instructions. **c)** Relative weighting of morphology was slightly reduced following the control instructions (top), but considerably reduced following the intervention instructions (middle). The bottom point shows the difference between the change seen in the two conditions.

Data Screening

All participants' responses were again at least somewhat influenced by either the luminance or the morphological cues, and so no data were excluded for this reason. Thirty-four of the 12,960 trials were dropped due to connectivity issues, and we excluded 167 trials with RTs greater than 4,650

msec (ten times the median), and 292 additional trials with RTs more than 3 standard deviations above that participant's mean. RTs were measured from the onset of the response prompt.

Reduction of Bias

Figure 4 shows posterior parameter estimates for the logistic mixed model fit to participants' responses. Complete posterior parameter estimates can be found in the supplementary materials. For all parameters reported, more than 99% of the posterior density fell on the same side of 0. Therefore, we can be extremely confident in the direction of these parameters. Prior to the instructions, participants' responses were again influenced by both luminance, B = [1.97, 1.74], control and intervention conditions respectively, and morphology, B = [1.06, 0.85], Figure 4a. We calculated the relative weighting of morphology, defined as the ratio of these regression weights and denoted β ((Figure 4b). The relative influence of morphology was reduced following the instructions in both the control condition, B = 0.54 (pre) to B = 0.45 (post), δ = -0.09, and considerably more reduced following the intervention instructions, B = 0.49 (pre) to B = 0.18 (post), δ = -0.31. Crucially, the reduction in relative influence of morphology following the intervention instructions clearly exceeded that seen following the control instructions (Figure 4c), Δ (difference in differences) = -0.22, 95% Bayesian Credible Interval = [-0.06, -0.39], $P(\Delta < 0) = .996$.

Analysis of confidence ratings and response times did not reveal any additional effects of interest. These analyses are reported in the supplementary materials.

Early or Late Bias Reduction

Figure 5a shows the average cursor position over time for a subset of stimuli, before and after instructions and separately for the control and intervention conditions. Figure 5b shows the influence of morphology and luminance on these positions, estimated from regression models fit to each 20 msec interval. Figure 5c shows planned comparisons contrasting the influence of morphology before and after the instructions were presented, along with the difference between the contrast seen in the intervention condition and that seen in the control condition. In both conditions, the influence of morphology, captured by the difference between the no-conflict and conflict stimuli here, was reduced following the instructions. Following the control instructions, the influence of morphology was significantly reduced from 1,700 msec after the onset of the target face onwards. Following the intervention instructions, there was a significant reduction from 1,000 msec onwards. Crucially, the reduction seen following the intervention instructions was significantly greater than that seen following the control instructions from 620 msec onwards. Recall that in Experiment 1

morphology began to influence participants' cursor movements from 460 msec onwards. Thus, the interventions reduced the influence of morphology from close to the time it first arises.

This conclusion was further supported by additional analysis of participants' responses and response latencies. First, we calculated participants' mean movement initiation and response times before and after the instructions, in each condition (see Table 1). If participants overcame their bias by taking longer to respond, we would expect a greater change in latencies for the intervention participants than for the control participants. Two-way ANOVAs revealed that participants were significantly faster to initiate the movements $F(1, 86)$[1] $= 19.130$, $p < .001$, $\eta^2 = 0.05$, and to respond, $F(1, 88) = 7.620$, $p = .007$, $\eta^2 = 0.004$. Importantly, however, there was no condition x stage interaction for initiation times, $F(1, 86) = 0.919$, $p = .0340$, $\eta^2 < 0.001$, $BF_{10} = 0.245$, or for response times, $F(1, 88) = 0.387$, $p = .535$, $\eta^2 < 0.001$, $BF_{10} = 0.234$. We also tested whether participants whose responses were more affected by the intervention – those with a higher Δ parameter in their analysis of responses – slowed down their responses more than those who were less affected. However, there was no significant correlation between changes in response times and changes in bias, $r = -0.08$, $df=111$, $p = .373$, $BF_{01} = 3.16$.

Condition	Stage	Initiation time (msec)	Response time (msec)	Confidence (%)
Control	Pre	-459 (293)	554 (233)	85 (8)
Control	Post	-539 (285)	533 (261)	86 (8)
Intervention	Pre	-331 (288)	665 (225)	87 (6)
Intervention	Post	-378 (292)	630 (220)	87 (6)

Table 1. Mean (SD) movement initiation times, response times, and confidence ratings, by condition, in Experiment 2, collapsing across the two intervention conditions. Initiation and response times are relative to the onset of the response prompt. Participants were significantly faster to initiate movements and to respond in the second half of the experiment, but this did not interact with condition.

1 Two participants were excluded from this analysis due to difficulties in calculating their movement initiation times.

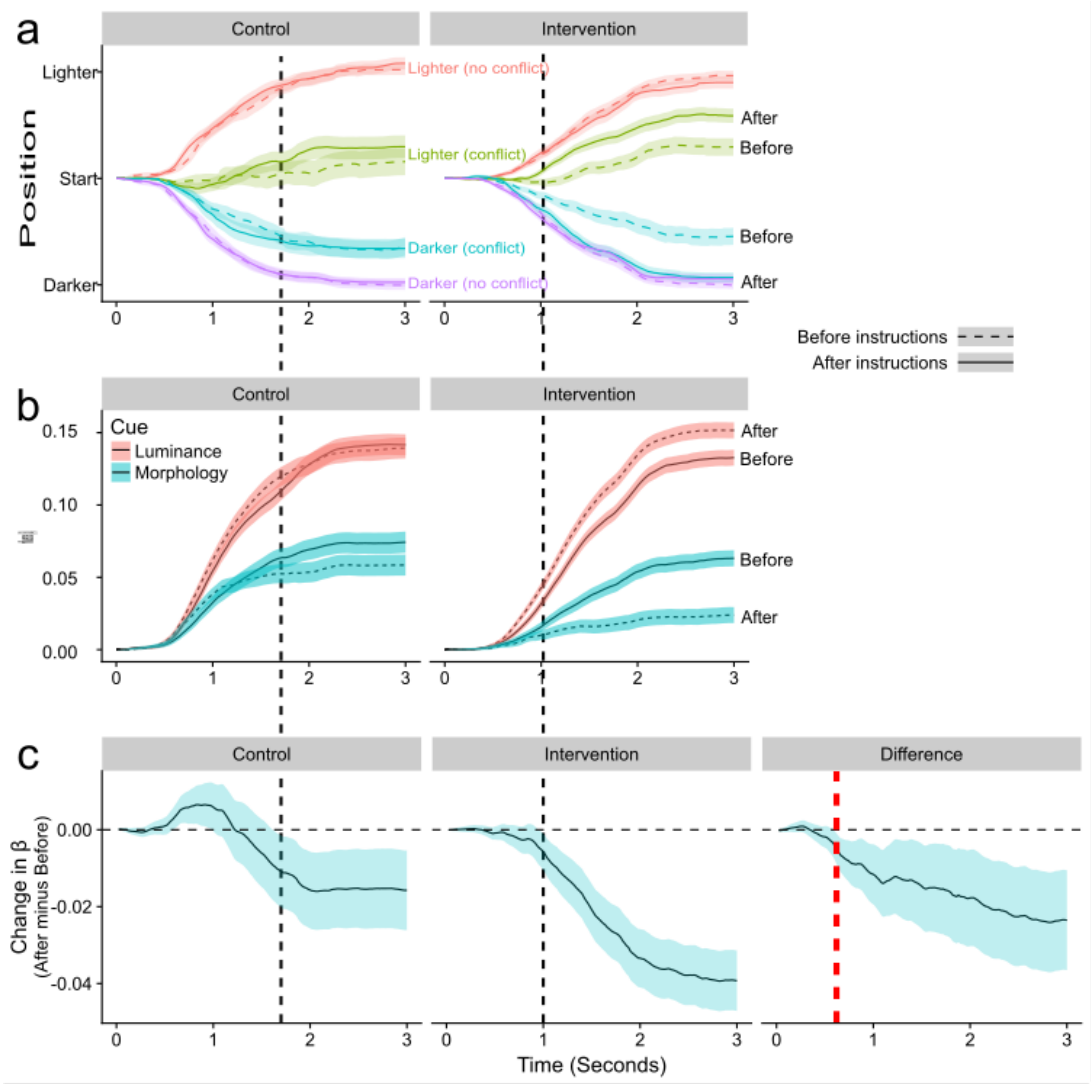

Figure 5 Effect of instructions on cursor trajectories, Experiment 2. a) Mean x-axis position of the mouse cursor over time, for the subset of trial trials where the target face as two levels darker or lighter than the reference, and the morphological cues either agreed or conflicted with the luminance. Solid lines show data from before instructions were presented. Dashed lines show data from after. b) Regression weights for the influence of morphology on cursor positions over time, before and after each type of instructions. c) Planned comparisons contrasting the influence of morphology before and after instructions for the control instructions (left), intervention instructions (centre), and the difference between the the two (right). Vertical black dashed lines show times from which the effect of morphology was significantly reduced: from 1,700 msec following control instructions, and from 1,000 msec following intervention instructions. Dashed red line shows point at 620 msec from which post-intervention change significantly exceeds post-control change.

Discussion

These results show, first, that the race-lightness effect can be reduced by making participants aware of their bias and asking them to try to avoid it. Both telling participants they were influenced by 'racial stereotypes' and telling them they were influenced by 'facial features' reduced the influence of morphology bias, compared to our control instructions, which did not refer to a bias. We did not find any notable differences between the two sets of instructions used to describe the bias. Unexpectedly, we also found that bias was somewhat reduced by our control instructions, although to a lesser degree than following the interventions. Importantly, all of our analyses accounted for this apparently spontaneous decrease in bias and showed effects above and beyond what was seen in the control condition.

Second, our mouse trajectory data show that bias was reduced even early in the decision process. Furthermore, participants did not appear to simply produce less biased responses by responding more slowly, so that an initially-biased response could be withheld. Together, these results indicate that asking participants to avoid the race-lightness effect alters either how participants initially perceive faces, or participants' immediate behavioural responses to them. In either case, these results show that the race lightness effect is to some degree subject to voluntary control.

General Discussion

The perceived lightness of another person's skin is biased by that person's facial features. We sought to outline the mechanisms that underlie this bias. First, we found tat that the bias arises early: morphological cues affected our participants' actions just as early as true luminance cues did. Second, we found that participants do not detect when this bias leads them to commit errors in a judgement task. Third, we found that the bias can be reduced, from close to its onset, simply by drawing participants' attention to its existence and asking them to try to avoid it.

Having outlined these processing characteristics, we now consider some of the implications of our findings for current theories in social cognition and perception. We related them here to two debates, one regarding the origin of the bias, and the other regarding its characterisation as an automatic process.

Beyond Top-Down/Bottom-Up

It is common to ask whether phenomena such as the race-lightness effect are 'top-down', cognitive, and conceptual, or 'bottom-up', perceptual, and automatic (Firestone & Scholl, 2014; Levin & Banaji, 2006; Deroy, 2013). We show that the race-lightness effect occurs quickly, and without metacognitive awareness. It would be tempting to conclude from this that the effect is bottom-up. However, we do not believe this interpretation is warranted, for a number of reasons. First, some top-down processes have been shown to occur rapidly, and unconsciously, while some classic bottom-up processes are nevertheless influenced by explicit expectations. Second, there is a growing consensus that few are exclusively bottom-up or top-down. Instead, the emerging view, both in social cognition (Freeman et al., 2012) and the cognitive sciences more broadly (Lupyan, 2015), is that there is a constant exchange of information between perceptual, cognitive, and motor processes, and, biologically, between levels of the cortical hierarchy. It should also be noted that it is notoriously difficult to assess the precise timing of mental processes using behavioural methods. Indeed, the time at which the bias emerges in our mouse cursor data, around 460 msec, is beyond the range of what can be thought of as early visual processing, given that ERP results suggest face categorisation can occur after as little as 200 msec (Liu, Harris, & Kanwisher, 2002). We would note, however, that our estimates reflect an upper bound on how long these processes can take – the point at which they manifest in overt motor movements. Future research might use EEG measures of visual responses to luminance and morphology, face categorisation, and motor preparation to more precisely explore how the race-lightness effect unfolds over time. Similarly, the asynchronies between the onset of the first and second faces here were somewhat longer than those used in other

behavioural studies of racial biases. However, participants could not begin to make a decision until the onset of the second face, and they saw the morphological cues at exactly the same time as they saw the key luminance cues. As such, it is unlikely that these long asynchronies would make it possible for participants to inhibit the morphological information.

Local Differences in Luminance

It has been suggested that the race lightness effect may be due to local differences in luminance (e.g. Firestone & Scholl, 2016; Laeng et al, 2018). Real and computer-generated face images are not uniformly bright (Figure 6). Instead, the centre (around the eyes and nose) tends to be brighter than the periphery, and this effect is greater in faces with African features. This means that when these images are normalised, although they have the same average luminance, faces with African features tend to be darker around the periphery and lighter in the centre than faces with European features. Figure 6 a and b shows this difference across our stimuli. The important point however here, for the current experiment, rests with the difference (Figure 6c) : since participants tend to foveate the centre of these images (e.g. Laeng et al, 2018), and the centre is lighter for faces with typically African features, local differences in luminance alone cannot explain the race-lightness effect in our results (it would, in fact, predict an effect in the opposite direction).

Laeng et al (2018) also noted that participants showed greater pupil constriction when viewing faces with African features. They suggested that this might provide a paradoxical explanation for the race-lightness effect. Since luminance-normalised images of faces with African features are brighter in the regions where the fovea is focused, fixating these images causes the pupil to constrict, allowing in less light overall, and so these images are perceived as darker than they really are. However, the pupillary light reflex takes a a few hundred milliseconds to occur, and is slower for weaker changes in luminance (Ellis, 1981). The differences in pupil size found by Laeng et al (2018) were only significant almost two seconds after stimulus onset. In contrast, our mouse tracking analyses revealed significant effects of morphology approximately half a second after the onset of the second face. Therefore, pupillary light reflexes also do not suffice to explain the race-lightness effect documented here.

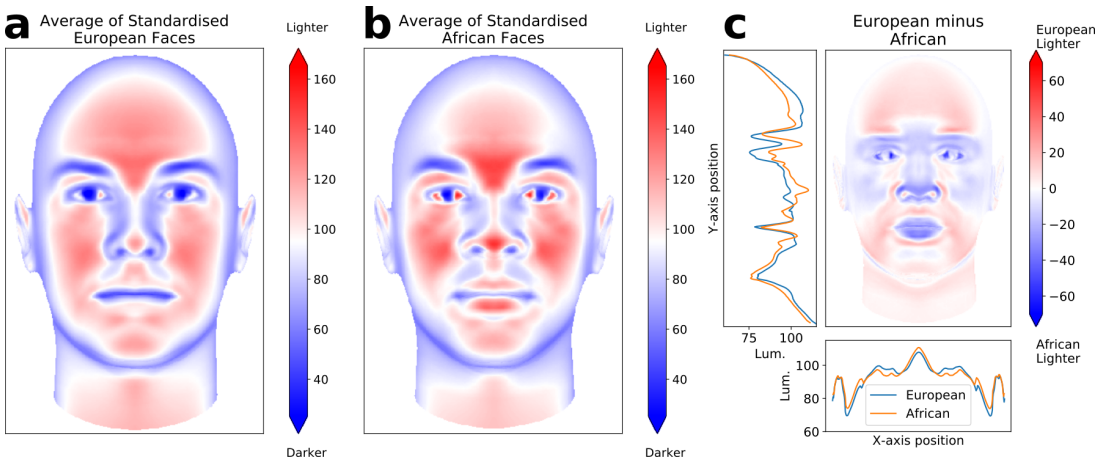

Figure 6. Local differences in brightness for luminance-normalised faces. **A.** Mean luminance per pixel, averaging across the eight faces with European features. Pixels of average luminance are shown in white, lighter than average in red, darker in blue. **B.** Mean luminance averaging across the eight faces with African features. **C.** Differences in mean luminance per pixel. Although all normalised faces have the same mean luminance when averaged across pixels, faces with African features are lighter around the eyes and nose, and darker in the periphery. Note scale differs from other panels.

Confidence and Awareness

It is also worthwhile to consider what can be inferred from our measurement of post-decision confidence ratings, and our finding that they do not dissociate from participants' responses. A cautious claim we can make from this is that the bias is at least persistent: the influence of morphological cues on participants' judgements is not a short-lived response bias, which can be easily spotted and corrected within a few moments of making the initial decision. In contrast, the best-known bias in social perception, where innocuous items held by African American individuals are misperceived as weapons does not persist beyond the first few seconds, and participants are easily able to spot their errors and reduce their subsequent confidence ratings accordingly (Payne et al., 2005). Going further, our results are consistent with the claim that participants do not have metacognitive access to the actual luminance of the faces they saw. This indicates that participants do combine these cues early on in processing, and do not or cannot separate them further downstream.

One potential issue with this interpretation is that participants might have access to the true luminance of the stimuli and may even be consciously aware that they have made use of morphological cues to make their decision, but they believe that doing so is entirely rational and consistent with the demands of the task. In other words, although participants were asked to make a judgement about which face was lighter, they may have instead decided to report and provide confidence for a judgement about which face looked more European. We cannot rule out this

possibility based on our data alone. However, we note that the race-lightness effect has been demonstrated using a range of paradigms, including direct colour-matching procedures (Levin & Banaji, 2006; Firestone & Scholl, 2014). Given that this bias is well-documented in cases where it cannot be due to participants answering the wrong question, it is unlikely that this explanation accounts for the bias seen here.

Automaticity

Another question that we may ask, given these results, is whether the race-lightness effect can be said to be automatic. There is a long-standing distinction in social and cognitive psychology between processes that are automatic and those that are controlled (Chaiken & Trope, 1999; Evans, 2008). The processing distinctions we discuss here – between fast and slow mechanisms, short- and long-lasting effects, the absence or presence of conscious awareness, being penetrable or impervious to high-level intentions – have in the past been included in lists of the features that distinguish automatic and controlled processing (Evans & Stanovich, 2013). Our results, however, do not support such a clear distinction. The race-lightness effect appears to show some of the classic hallmarks of automaticity, in that it arises relatively quickly, and apparently without participants being aware of the cue integration process. However, it is also sensitive to high level goals (see Luque, Vadillo, Lopez, Alonso, & Shanks, 2017, for recent analogous discussion in a different domain). Results from many studies however here concur in suggesting that no cognitive phenomena are completely automatic, or completely controlled (Jacoby, 1991; Payne, Jacoby, & Lambert, 2004). Our findings are consistent with this idea that there is continuum between automatic and controlled phenomena, and dissociations between the proposed features of automaticity.

What our results certainly do is to further our understanding of the awareness and control of biases and illusions arising from cue integration. Research on cue integration across sensory modalities has focused on how the combination of information from multiple senses can lead to perceptual biases, such as spatial ventriloquism, and illusory percepts, such as the McGurk effect or the double-flash illusion (Deroy et al., 2016). Previous work suggests that participants are less confident in percepts that come from combining two incongruent perceptual sources than those that come from sources that are congruent (Hillis, Ernst, Banks, & Landy, 2002; White et al., 2014). However, this finding has only been investigated in cases where the information comes from different sensory modalities (e.g. vision and audition), but not from different cues within a single modality (Hillis et al., 2002), as is the case with luminance and morphology in the race-lightness effect. Importantly, our results contrast with findings obtained in multisensory perception and

suggest that cue integration within vision alone is not accessible to metacognition and not reflected in participants' confidence ratings.

Conclusion

To conclude, these results bring an important qualification to the claim that 'racial' biases are driven by fast, reflexive processes, and that they are therefore beyond one's control or responsibility. Participants were not aware of their errors, but those errors could nonetheless be reduced by bringing them to people's attention and asking them to try avoid them.

References

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience & Biobehavioral Reviews*, *81*, 16–28. https://doi.org/10.1016/j.neubiorev.2017.02.012

Bruner, J. S. (1973). Beyond the information given: Studies in the psychology of knowing. W W Norton & Company Incorporated.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1–28.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A probabilistic programming language. Journal of Statistical Software, 20, 1–37.

Chaiken, S., & Trope, Y. (1999). Dual-process theories in social psychology. Guilford Press.

Deroy, O. (2013). Object-sensitivity versus cognitive penetrability of perception. Philosophical studies, 162(1), 87-107.

Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in multisensory perception. Trends in Cognitive Sciences, 20(10), 736–747. https://doi.org/10.1016/j.tics.2016.08.006

Ellis, C. J. (1981). The pupillary light reflex in normal subjects. *British Journal of Ophthalmology*, *65*(11), 754–759. https://doi.org/10.1136/bjo.65.11.754

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. Visual Cognition, 2(2–3), 101–118. https://doi.org/10.1080/13506289508401726

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the Debate. Perspectives on Psychological Science, 8(3), 223–241. https://doi.org/10.1177/1745691612460685

Firestone, C., & Scholl, B. J. (2014). Can you experience 'top-down' effects on perception?: The case of race categories and perceived lightness. Psychonomic Bulletin & Review, 22(3), 694–700. https://doi.org/10.3758/s13423-014-0711-5

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. Behavioral and Brain Sciences, 39. https://doi.org/10.1017/S0140525X15000965

Fleming, S. M., & Frith, C. D. (2014). The cognitive neuroscience of metacognition. Springer Science & Business Media.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8. https://doi.org/10.3389/fnhum.2014.00443

Fodor, J. (1983). The modularity of mind: An Essay on faculty psychology. MIT Press.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. Psychological Science, 20(10), 1183–1188.

Freeman, J. B., & Ambady, N. (2011a). A dynamic interactive theory of person construal. Psychological Review, 118(2), 247–279. https://doi.org/10.1037/a0022327

Freeman, J. B., & Ambady, N. (2011b). When two become one: Temporally dynamic integration of the face and voice. Journal of Experimental Social Psychology, 47(1), 259–263. https://doi.org/10.1016/j.jesp.2010.08.018

Freeman, J. B., Johnson, K., Adams, R. J., & Ambady, N. (2012). The social-sensory interface: category interactions in person perception. Frontiers in Integrative Neuroscience, 6. https://doi.org/10.3389/fnint.2012.00081

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Greenwald, A. G., Mcghee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition : The implicit association test. Journal of Personality and Social Psychology, 74(6), 1464–1480.

Greenwald, Anthony G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Gregory, R. L. (1968). Perceptual illusions and brain models. Proceedings of the Royal Society of London. Series B, Biological Sciences, 171(1024), 279–296.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. Science, 298(5598), 1627–1630.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. Journal of Memory and Language, 30(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. Annual. Review of Psychology., 55, 271–304.

Laeng, B., Kiambarua, K. G., Hagen, T., Bochynska, A., Lubell, J., Suzuki, H., & Okubo, M. (2018). The 'face race lightness illusion': An effect of the eyes and pupils? *PLOS ONE*, *13*(8), e0201603. https://doi.org/10.1371/journal.pone.0201603

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. Social and Personality Psychology Compass, 7(5), 315–330. https://doi.org/10.1111/spc3.12023

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

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: The role of race categories. Journal of Experimental Psychology: General, 135(4), 501–512. https://doi.org/10.1037/0096-3445.135.4.501

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. Nature Neuroscience, 5(9), 910–916. https://doi.org/10.1038/nn909

Lupyan, G. (2015). Cognitive Penetrability of Perception in the Age of Prediction: Predictive Systems are Penetrable Systems. Review of Philosophy and Psychology, 6(4), 547–569. https://doi.org/10.1007/s13164-015-0253-4

Luque, D., Vadillo, M. A., Lopez, F. J., Alonso, R., & Shanks, D. R. (2017). Testing the controllability of contextual cuing of visual search. Scientific Reports, 7, 39645. https://doi.org/10.1038/srep39645

MacLin, O. H., & Malpass, R. S. (2003). The ambiguous-race face illusion. Perception, 32(2), 249–252. https://doi.org/10.1068/p5046

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. Annual Review of Psychology, 51(1), 93–120.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746–748. https://doi.org/10.1038/264746a0

Morgan, M. J., Hole, G. J., & Glennerster, A. (1990). Biases and sensitivities in geometrical illusions. Vision Research, 30(11), 1793–1810. https://doi.org/10.1016/0042-6989(90)90160-M

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (2004). Memory monitoring and the control of stereotype distortion. Journal of Experimental Social Psychology, 40(1), 52–64. https://doi.org/10.1016/S0022-1031(03)00069-6

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. Journal of Experimental Social Psychology, 41(1), 36–47. https://doi.org/10.1016/j.jesp.2004.05.001

Rausch, M., Müller, H. J., & Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. Consciousness and Cognition, 35, 192–205.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. Trends in Cognitive Sciences, 18(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Stocker, A., & Simoncelli, E. P. (2007). A Bayesian model of conditioned perception. In Advances in neural information processing systems (pp. 1409–1416).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643–662. https://doi.org/10.1037/h0054651

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. Cognition, 150, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. Proceedings of the National Academy of Sciences, 111(45), 16214–16218. https://doi.org/10.1073/pnas.1403619111

White, T. P., Wigton, R. L., Joyce, D. W., Bobin, T., Ferragamo, C., Wasim, N., … Shergill, S. S. (2014). Eluding the illusion? Schizophrenia, dopamine and the McGurk effect. Frontiers in Human Neuroscience, 8, 565. https://doi.org/10.3389/fnhum.2014.00565

Zeki, S. (1993). Vision of the brain. Blackwell Publishing, Inc

Author Contributions

All authors developed the study concept, and contributed to the study designs. E.T. generated the stimuli, and programmed the task. E.T. performed the data analysis, and all authors interpreted the results. E.T. and O.D. drafted the manuscript, with critical revisions by M.T.F. All authors approved the final version of the manuscript for submission.

Competing Interests

None

The authors declare no competing financial interests.