Predictions do not entail cognitive penetration:
Racial bias and predictive coding models of perception

Ophelia Deroy
Chair, Philosophy of Mind
Faculty of Philosophy & Munich Centre for Neuroscience
Ludwig-Maximilians University Munich

ophelia.deroy@lmu.de

TO BE INSERTED : Versions of the present paper were presented at the 40th International Wittgenstein Symposium in Kirchberg, the PLM conference in Bochum, the ECAP in Munich, and at departmental seminars in Salzburg and Magdeburg. Thanks to the organisers and audience of these events, for their very insightful and helpful comments. Particular thanks to Christopher Gauker, Sascha Fink, Christoph Limbeck, Albert Newen and Jesse Prinz for comments, and to Merle Fairhurst and Eoin Travers for their contribution to the empirical and theoretical arguments exposed in this paper.

1. Introduction

Racial perceptual biases are, rightly so, among the most discussed in the literature. Not only do they confirm that our perception of the world is not always accurate, but their inaccuracy comes with worrying ethical consequences. This concern starts with the suggestion that racial categories - whose use and even mention are highly questionable - have a pervasive influence on our mental lives, including the most basic perceptual processes. Among the most concerning biases to be documented in perception is the tendency to categorise an ambiguous object as a gun when it is carried by a black man, and as a mere phone or tool when carried by a white man (Correll et al., 2002; Payne, 2006; see also Ofan et al, 2011 for a priming study). Another, no less sadly significant bias, is the ‘cross-race’ recognition deficit (Levin, 2000; Hugenberg et al., 2007). This well-documented effect shows that individuals tend to perceive members of another ethnic group as more similar to one another than members of their own - a bias which has also impact on re-identification of criminals in court cases (Sporer, 2001).

However important to the ethicists, these biases are not the ones that attract most attention from philosophers of perception. One bias, documented by Levin & Banaj in 2006, has been the object almost of a monopolistic interest from experts in perception: This so-called ‘race-lightness’ bias shows that the perceived lightness of skin is influenced by the morphology of the face it belongs to, depending on whether it corresponds to a typically African or a typically European morphology. Though human faces present a high degree of distinction, there is evidence that there morphological and skin tone differences correlate between humans ¹, making it likely that this correlation is learned by human perceivers. The authors’ hypothesis was to test whether this prior knowledge about the reflectance of faces also affects their perceived lightness.

In the most telling of their experiment, participants were asked to adjust the luminance of a test face to match the luminance of a target face. The target face could present either a typically African or typically European morphology, and vary in luminance (five values). The response face was also either of the same morphology or a different one, to counter-balance the design. What the results show then is that faces with the same objective luminance were not perceived as equally light or dark depending on the facial features of the target: Faces with typically African morphology were consistently matched to a darker shade than faces with typically European morphology, though their mean objective luminance was similar.

In another experiment, participants were presented with the same face, accompanied either by the label ‘Black’ or ‘White’, and similar biases were observed: Here, both luminance and morphology being equal, a face was perceived as darker when labeled as ‘Black’ or ‘White’.

The reason why this bias is so important is that, by contrast with the gun/tool bias, it is not about a difficult situation of perception. The stimuli are presented without time constraints, and the response is also possibly slow. By contrast with the cross-race recognition deficit, it is clearly about perception, and not memory. More crucially, this bias is also a good case of non-motivated perception. Research has indeed shown that the responses in the shooter’s task were largely driven by fear responses (Azevedo et al., 2017). Last but not least, the bias here affects a continuous variable (the judgement of brightness) and can itself manipulated parametrically (by manipulating the morphological features through morphing).

The philosophical challenge arises then from this perceptual manifestation: as mentioned by Firestone and Scholl, (2016), “it is, to our knowledge, the only purported top-down effect on perception that readers can experience for themselves simply by looking at the stimuli.” While most experiments on top-down influences show effects on reports or judgements, which remain problematic, this effect shows a clear influence on perceptual experience (see Deroy, 2012 for discussion). The influence of thought, or its conceptual constituents, on perceptual experience is at the core of what philosophers have in mind when discussing cognitive penetration of perception (e.g., MacPherson, 2012, Siegel, 2012, Newen & Vetter, 2017) though it is certainly not the only version of cognitive penetration that is discussed (e.g., Pylyshyn, 1999, Raftopoulos, 2001, Zeimbekis & Raftopolous, 2015 for discussion).

What’s more, because the effect is not timed, nor reflected a special attentional scanning of the scene, attentional explanations more easily ruled out.

However, in this paper, I want to show that this interpretation is problematic. What’s more, framing this as a case of cognitive penetration is not helpful. I want to suggest a better problem that this case raises.

2. Three issues with the cognitive penetration interpretation of the bias

2.1. What is the source of influence?

Cognitive penetration accounts need to provide a plausible candidate for the source of the cognitive influence. However, it is really unclear what that can be. The strength of the effect does not correlate with explicit racist attitudes (tested). So can we identify other explicit beliefs? The challenge here to is account for the graded character of the effect, as well
as its probabilistic nature: Do people believe that “All individuals with African facial features have
darker skin-complexion than individuals with Caucasian facial features.”? That “Most individuals
with African facial features have darker skin-complexion than individuals with Caucasian facial
features”, or that “The more pronounced the African facial features, the darker the skin; the more
pronounced the Caucasian features, the lighter the skin.”.

The difficulty, not to say the failure, of identifying a good explicit cognitive candidate pushes one
toward other, implicit, candidates.

Empirical evidence in the domain of other racial biases points at the role of implicit and associative
biases (e.g., IAT, Greenwald et al., 2015 for recent review) - but whether they are conceptual or not
remains debated. This point leads to a second challenge.

2.2. Is the influence genuinely cognitive?

Besides the problematic role of implicit associations, the literature on biases also points at the role
of intermediate levels of categorization, which are pretty different from what we mean by the
contents of thoughts. Those categories are inherently probabilistic, continuously and quickly
updated through exposure, and not consciously accessible (Rigoli et al., 2017, for review).

Several people (including myself, see Deroy, 2013) have highlighted the fact that not all top-down
influences should count as cognitive in some of the discussed cases of ‘cognitive penetration’. Of
course, this recommendation is only to be taken seriously if there is a good account of the non-
cognitive candidates. As it happens in the race-lightness bias, there is

2.3. Is this top-down at all?

An independent challenge, however, has arisen in the case of the race-lightness bias: Do we need to
resort to top-down explanations at all? In a recent study, Firestone and Sholl (2014) used blurred
faces instead of clear ones. Blurry African face still perceived as comparatively darker than blurry
European faces by 70% participants, and this despite the absence of awareness of racial differences
between faces. In this case, it seems that the distribution of darker and lighter pixels in the blurred
versions of the faces drives the effect, turning the bias actually into a reflection of a true difference
between the stimuli. As the authors conclude, if differences in the stimuli are sufficient to explain
the effect, there is no need to appeal to top-down influences, cognitive or not. The results of this
study have been challenged, and discussed, but this is actually beside the point. First, from the
evidence provided by the authors, one can exclude that there is a conscious recognition of the
differences between the two faces, but not that there is a non-conscious one, triggering eventually a
high-level conceptual association, or racial categories which exert an influence on the perception of
skin lightness. The results are still compatible then with a top-down influence whose source remains
unconscious. There is a second issue, this time independent of some alternative explanation of the
effects. For even if one grants the conclusion that differences in the distribution of luminance in the
blurred stimuli drive the difference in perception, this conclusion does not necessarily generalise to
non-blurred cases. In other words, if one does not have sufficient information to categorise or
recognize African faces as African, and European faces as European, when the stimuli are blurred,
this information becomes available when the faces are not blurred. Then, it is perfectly possible that
the low-level differences do not play much role, and that top-down influences from the categories
or concepts triggered by the stimuli do drive the effect. The role of bottom-up differences with
blurred faces, when ethnic categorisation is not available, does not preclude that, when ethnic categorisation is possible, with non-blurry faces, there is no role for top-down influences. Indeed, it is even possible to consider that the two explanations are then still valid: The effect in the non-blurry case could be due to a combination of both sensory differences and top-down influences.

2.4. Intermediate summary

The Race-lightness is a robust example of perceptual bias, where differences in the final percept seem to be driven by prior expectations, rather than by differences in the world. Though the effect is most interpreted as an instance of cognitive penetration, this interpretation, however, is problematic in at least three aspects:

First, it is unclear which cognitive state can be cited as a good candidate for the cognitive influence; Second, it is not fully justified why cognitive candidates should be privileged over other top-down non-cognitive influences; Finally, it is not even fully established that top-down influences are at all needed as an explanation.

Each of these problems points at different weaknesses of the account. Some, it could be argued, are conditioned on some empirical evidence still missing, that could eventually be provided by cognitive neuroscience studies. Others, however, seem to be more fundamentally problematic, such as to establish whether unconscious non-cognitive candidates operate or not. There is, undoubtedly, no way to predict what future experimental designs will be able to do, but the primary issue for ‘strict’ cognitive penetration accounts is that they face a competitive explanation, which can overcome all these difficulties.

3. Predictive accounts, and the end of the “cognition influences perception” debate?

3.1. Why predictive accounts count as best available explanation for a bias like race-lightness.

Predictive coding accounts, which are being defended by researchers like Friston, in the neurosciences, and Clark, Hohwy, Lupyan in philosophy, provide a way to accommodate all the previous evidence and lack of determinate answer regarding the cause of the race-lightness bias. As perceptual processing is redefined in such accounts as a combination of top-down and bottom-up factors, and several layers of top-down influences, ranging from the high-level cognitive influences, with explicit beliefs and propositional attitudes supported by language at the top, to low level, usually precise predictions at the bottom. At each level, predictions are matched to the incoming sensory signal, with only the error (the mismatch between predictions and input) being transferred to the higher level. This model is supposed to explain both the stability of perception, its flexibility in front of new evidence.

In the case of the race-lightness bias, predictive accounts can accommodate both some differences in the stimuli and multiple sources of top-down influences, now redefined as predictions. The higher-level predictions, being the most abstract and general, cascade down to more and more specific predictions about basic sensory features, such as, eventually, brightness processing.

---

This interpretation overcomes the dilemmas in which the cognitive penetration interpretation was drowning. It raises, however, one problem - at least for philosophers: It seems to dissolve all problem of thought influencing perception, leaving us to understand only how different levels of predictions and error signals explain the perceived bias. More importantly, predictive accounts are usually considered as initiating ‘a collapse of perception and cognition’ (Lupyan, 2015). This diagnosis is echoed by philosophers, such as Andy Clark when he states that predictive coding “makes the lines between perception and cognition fuzzy … In place of any real distinction between perception and belief we now get variable differences in the mixture of top-down and bottom-up influence …” (Clark, 2016) and by neuroscientists supporting predictive accounts, such as Karl Friston and Chris Frith, who admit that “within this framework, there is no qualitative distinction between perception and belief, since both involve making inferences about the state of the world on the basis of evidence ” (Frith & Friston, 2013).

This diagnosis, I want to argue, is mistaken: The cognition/perception divide persists in predictive coding accounts, and therefore, the question of whether thought influences perception still makes sense for a proponent of those accounts.

3.2. The cognition/perception divide in predictive accounts

I don’t want to deny that predictive coding accounts raise a challenge for the way perception works. Right from the earliest level of processing, what is processed is the mismatch between a prediction coming from a higher level in the hierarchy, and the incoming sensory signal. Still, besides this significant redefinition, every level, in this respect, receives some sensory signal from the level below. Not all levels, however, receive predictions from the level above. By necessity, predictions must start somewhere, and this upper level of the hierarchy is then the one which predicts but is not predicted by anything else. What I suggest is to call this “thought.”

Importantly, the almost trivial observation that the upper level is not predicted by anything else means that, instead of challenging the cognition/perception distinction, predictive coding accounts provide a substantial characterisation of thought.

Obviously, thought gets updated by the flow of error signals coming from the immediately inferior level, which we can generally equate to the perception of objects and scenes. Depending on what one believes perception to be (i.e., akin to or different from beliefs) then the characterisation of the level below thought might change, but this point is orthogonal to our current question. We can expect then that, exposing people to a large number of faces where morphology is not correlated with skin luminance, or where it presents an inverse relation to the present one (i.e., African faces appearing with lighter facial luminance, and European faces with darker facial luminance) will slowly correct the predictions. It is still not obvious how to measure how this change will impact thought. To do so, we have to move to a positive, rather than negative characterisation of thought in predictive coding accounts. Instead of merely noticing that the upper level, i.e., thought, is not predicted by anything else, we can posit that it can be updated by something else than perceptual error signals: Other thoughts linguistically communicated by others. By contrast with perceptual evidence, which should lead to a slow updating of the predictions, linguistic statements can change one’s thought through single exposure (though quantifying this is a delicate matter). The prediction about racial biases now becomes more substantial: If higher-level cognition is involved in predictions, which cascade down to influence lower-level brightness processing, then by initiating a change at this level, by providing the perceiver with a linguistic testimony and/or a set of new
instructions, the change should have repercussions all the way down, and change the effect. Because of the stability of lower-level predictions, and because of persisting differences in the initial stimuli, one would also expect a reduction, rather than a disappearance of the bias. This is indeed what is observed with participants who are told about their biases. Travers et al. (2017) ran a variant of the race-lightness effect, inspired by Levin & Banaji (2006), with 90 participants. The first 70 trials confirmed the existence of a robust bias, where morphology would influence the perceived luminance of faces. After those first trials, however, participants were divided in three groups: The first group was informed that they were relying on morphology in their judgements, and advised to do better in the next trials; the second group was informed that they were relying on racial stereotypes in their judgements, and also advised to do better in the next trials. The third group of participants acted as a control, and were just ask to take a rest and to try and do better in the next trials. The results show that the first two groups, whose thoughts and goals had been reset by instructions, showed a significant reduction of the perceptual bias in the next set of trials. Importantly, the time-course of the responses suggest that the reduction of the bias is not due to an inhibition of responses, but to a genuine lowering of the weight given to morphology. In other words, the experiment successfully demonstrates that a change in one’s thoughts (or higher-order level cognition) triggers a change in the race-lightness effect, in a way which is congruent with the predictive coding accounts and the idea of a possible, yet indirect influence of thought on perception.

4. Conclusions

I have here argued for two things. First, perceptual biases, instead of fuelling debates about the direct influence of thought on perceptual experience and processing, should be interpreted in a more distributed way, as revealing an eventual, yet not total influence, of thought on perceptual processing and experience. In other words, the question we should be asking is not “Does thought (or higher-level cognition) influence perceptual experience and processing?” but “How much does thought influence perceptual experience and processing?”

Second, we should be testing the difference that thought can make to perception, rather than focusing only on interpreting effects. The right question to ask, in this respect, is not just “How much does thought influence perceptual experience and processing?” but “How much can a difference in thought trigger a change in perceptual experience and processing, other things being equal?” Here, I believe we start to have good evidence, in the case of the race-lightness effect, that a change at the higher level of cognition can trigger a change in perception.

Although I have framed the argument around one specific bias (the ‘race lightness bias’) it should be valid for others. The argument also rests on predictive accounts but is possibly detached from some of their assumptions (such as the fact that all types of processing in the brain are predictive and Bayesian). Though I am positively convinced that the predictive models are robust and enable us to approach the mind/brain better than other models, the arguments involved here do not require a specific form of realism or exclusivity toward such models. What matters still is to show that,

---

3 See also Orlandi (2015) for an in-depth discussion of this issue. Though she reaches a more radical conclusion than the present one, she also argues that vision does not need to count as cognitive in any sense in predictive or Bayesian models,
when working within such models, we are not asked to give up the distinction between perception and cognition, but asked to re-consider new hypotheses and ways to approach it.

References


