The focus of this paper will be implicit biases where a person reports non-racist attitudes but still behaves in a racist way. I will use the following characterization:

(*implicit bias*) A mental state M of subject S is an implicit bias if

(i) M causes behavior B,

(ii) B "involves a deviation from norms of fairness," (Frankish, 2016)

(iii) B looks like S believes that p,

(iv) S reports that she has an explicit belief that not-p.

The fact that implicit biases involve a kind of mismatch between (iii) and (iv) is often used to argue that implicit biases are unconscious (Gawronski, Hofmann, & Wilbur, 2006, p. 487). Despite its wide acceptance and its initial plausibility, the claim that implicit biases are unconscious has recently come under attack. Most importantly, (Hahn, Judd, Hirsh, & Blair, 2014) have shown that subjects can correctly predict their own IAT-results if asked to focus on their gut feelings. (Hahn et al., 2014) explain their results on the basis of the so-called 'APE model' (*Associative–Propositional Evaluation*) (Gawronski & Bodenhausen, 2006). According to this model, implicit attitudes are based on spontaneous affective reactions that can, due to their phenomenal character (they can be felt), be introspectively accessed.

Gawronski uses the following example to illustrate how implicit biases arise (Gawronski, 2012, 662). Imagine a subject who comes to believe the following three claims, where the first belief is due to an affective reaction towards African Americans:

(1) I dislike African Americans.
(2) African Americans are a disadvantaged group.
(3) Negative evaluations of disadvantaged groups are wrong.

The belief set containing beliefs (1)-(3) is *cognitively inconsistent* (Gawronski, 2012). Gawronski argues that cognitive consistency is reestablished by rejecting either of the beliefs. Implicit biases arise when the subject resolves the inconsistency by rejecting (1). The implicit bias occurs due to the fact that the negative affective reaction that led to first holding (1) persists even though (1) is rejected. Furthermore, Gawronksi argues that this is an entirely conscious process.

The APE model is promising because it can explain how subjects can be introspectively aware of their biases. Still, the APE model is problematic. First, since the APE model assumes that the reestablishment of cognitive consistency is a conscious process, it cannot account for the fact that not all subjects in Hahn et al.'s study could become aware of their biases, and second it cannot explain why people are surprised when realizing that they are biased. Third, it is highly implausibly that people can convince themselves to like someone, if they know that they do not like that person.

I will argue that the problems for the APE model can be avoided if one thinks of implicit biases in terms of repression. Based on Alexander Billon's account of repression in terms of Ned Block distinction between phenomenal- and access-consciousness, I will develop an improved model of repression that avoids several problems afflicting Billon's account while keeping the basic assumptions of the APE model. According to my analysis, repression consists of 7 steps:

Step 1: A feeling (a p-conscious state) is correctly but imprecisely categorized
Step 2: The subject has the explicit desire not to have the p-conscious state (under its imprecise categorization)
Step 3: Step 1 and 2 result in inner conflict that elicits negative feelings
Step 4: The subject impulsively avoids the conflict by either miscategorizing the feeling or by shifting attention
Step 5: Habitualization: the p-conscious state automatically leads to miscategorization or shift of attention
Step 6: Repression itself is an a-unconscious process
Step 7: Re-discovery of repressed states means to draw attention to the conflict and to categorize it correctly

Implicit biases are the consequence of repression in the following way: take a person who has an implicit negative bias towards people of color. This person, first, realizes that she has a negative feeling when interacting with people of color. She also has the explicit belief that if one wants to be a fair and good person one should not discriminate against groups on the basis of contingent properties of these groups. This person wants to be a good person. Hence, she has the explicit desire not to have negative feelings towards people of color. Her feeling and the desire induce an inner conflict with the preferred self-image and internalized social norms.

This conflict impulsively leads the person to either shift her attention away from the

feeling (by, e.g., thinking about something else) or the person miscategorizes the feeling (as, e.g., a bad mood, or a stomachache). Since she always does so in the presence of the p-conscious feeling (which is triggered by the interaction with people of color), at one point, the feeling starts to automatically activate the avoidance behaviors. The repression process itself occurs a-unconsciously because the person does not recognize that there is an emotion that needs to be repressed, or because she does not pay enough attention to her p-conscious state and to her behavior, or because she lacks the relevant concepts to categorize her behavior and the conflict in the right way.

We can avoid the problem afflicting the APE model. The approach suggested here requires the affective reaction only to be p-conscious, and a-conscious in a rather imprecise way. The subject does not have to make the evaluation "I don't like African-Americans" in order to, then, reject it. Furthermore, my account does not require that subjects can consciously convince themselves to like someone or something while knowing that they do not like that person or thing. The reason is that I do not take implicit biases to result from cognitive inconsistency that has to be resolved by rejecting a belief (that is based on a affective reaction in this case of implicit biases). Rather, implicit biases are due to the avoidance of a violated desire. Finally, based on the repression view of implicit biases one can make several interesting empirical predictions.

References

Frankish, K. (2016). Playing double: implicit bias, dual levels, and self-control. *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology2*, *I*, 1–22.

Gawronski, B. (2012). Back To the Future of Dissonance Theory: Cognitive Consistency As a Core Motive. *Social Cognition*, *30*(6), 652–668. http://doi.org/10.1521/soco.2012.30.6.652

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. http://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499. http://doi.org/10.1016/j.concog.2005.11.007

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology. General*, *143*(3), 1369–92. http://doi.org/10.1037/a0035028