« Animal In Mind: People, Cattle and Shared Nature on the African Savannah The Health Impact Fund: a better way to reward new medicines »

The Dark Dark Side of the Mind

by **Mahzarin R. Banaji**, Richard Clarke Cabot Professor of Social Ethics, Department of Psychology, Harvard University

It was a chilly fall night in Columbus Ohio and we had just returned, as we did almost every evening in graduate school, from a sumptuous dinner at Wendy's (the city was, after all, the testing ground for the highest echelon of fast food chains). As usual, we had done our part for food research by consuming a nontrivial number of strangely textured burgers and florescent milk shakes. We had returned, also as usual, to our desks located under the bleachers of the football stadium at Ohio State, where we would remain until late into the night. That evening, my friend and officemate Trish Devine told me that she was going to subliminally prime stereotypic concepts about Black Americans to see if such unconscious activation would affect reactions to subsequent racially ambiguous information.

"You've got to be kidding me," I thought, even though I'm sure I found something supportive to say. There was something illicit about the idea and it seemed to be such an invitation to trouble when perfectly good research on impression formation awaited. Trish could continue to counterbalance hundreds of sentences about John and Mary and test memory for them. She gets her data *and* nobody gets hurt. But now that the unthinkable had been suggested, how could one not wonder about what might happen, what the ramifications would be if it turned out that the result she had obviously imagined did indeed obtain.

Professor Mahzarin Banaji

To put my trepidation into context, remember that this was 1983. We still wore t-shirts that said "Help stamp out disco in our lifetime", we were certain that Ronald Reagan would lose the election, Michael Jackson's hair had not yet caught fire doing a Pepsi commercial. It is fair to say that what we now know to be such well-established facts about dissociations between conscious and unconscious feelings were hard to even imagine, let alone about race, because they involved aspects of the mind into which experiments hadn't yet been ushered. The methods to study such topics were available in dribs and drabs. In the 1970s, a few studies of amnesic patients showing dissociations between explicit and implicit memory existed (like the work by Warrington and Weiskrantz[1]); with ordinary people, not much evidence existed except the landmark work on semantic priming by David Myers[2] and Jim Neely's dissertation[3], but the most outrageous they had gotten in terms of the content, was to test if "bread" primed "butter". A few papers on subliminal perception were also available (did king prime queen?), but in 1983 we didn't trust such results any more than we trusted that a virus that caused AIDS would be discovered that year.

So what I can say with certainty is that I had no idea that my remarkable friend was about to make a breakthrough that would open a brand new gate to theorizing and experimentation about the double dark side of the mind. There were those who studied one of the darks — the mental unconscious where processes unfolded without conscious awareness, without conscious control and without intention or self-reflection. There were also those who had studied the other kind of dark — humans with high amounts of melanin in their epidermis. Trish would bring these two darks together in a tough test, and in her work the scientific method was used to unearth new knowledge about human groups and our representations of them.

The first experiments showed evidence for unconscious negative attitudes towards Black Americans in those of us who seemed unaware that they even had such attitudes. When the psychologist's probe bypassed the conscious mind, we all looked more alike — many more of us showed negative attitudes toward dark skinned people.[4] Now, it is more than 28 years later and mention of disco music only makes me smile nostalgically. I think Ronald Reagan did become president. And Michael Jackson, sadly, burned more than just his hair.

Today, as a science, we know a whole lot more about the nature of the implicit ways in which we think about social groups, and race in particular. After Trisha's experiments, I myself had joined the crew to try to figure it all out. It was impossible to do anything but.

Human beings have considerable capacity for self-deception[5] because the architecture and the traffic through the mind allows information to be dissociated: it allows us to both know and not know the same thing, to feel and not feel the same way about the same thing, to be aware and not aware of the motives that guide behavior toward ourselves and others. Psychologists and neuroscientists have found it profitable to think about the mind's fractures by relying less on the malign motives of bad social actors as the explanation, and more on elucidating the evolutionary presses that created the minds we have and the sociocultural and situational presses that exert influence on us more proximally.

Over the past 25 years, we have studied the hidden biases of good people, i.e., self-professed egalitarians. My colleagues and I include ourselves in this group, and some of our methods have allowed us to be subjects in our own experiments. This unusual opportunity has brought us closer to some answers that would otherwise have been difficult to accept if we hadn't had the direct experience we did. It also brought us answers more quickly because the clarity and strength of the evidence were hard to refute.

In my own lab in the mid 1990s, the opportunity to view implicit bias up close came in the form of a test of implicit cognition called the Implicit Association Test (IAT)[6]. Although it can be put to use in a wide variety of settings, the first experiments that were done with it focused on the attitudes we have toward a variety of social groups — female and male, Black and White, elderly and young — giving us a simple 2 X 2: those groups of which we are members (female, elderly) and those social opposites of which we are not (male, young); those groups that are relatively more preferred (female, young) and advantaged (male, young) and those that are less preferred (male, elderly) and less advantaged (female, elderly).

To grasp the processes about which we speak, it is best to take a test yourself by visiting **www.implicit.harvard.edu**, and given the focus here on race, you might try one of a few tests involving these human categories — i.e., tests that use as stimuli faces of black and white adults or children, or dark-skinned and light-skinned people more generally. You can take a test that measures the association of "good" and "bad" concepts such as "love", "peace", "friend" or "anger", "war", "devil" with the aforementioned social groups or one that measures the association of these groups with "weapons" and "harmless" objects. If you are like most White and Asian people who take these tests, you'll have a harder time associating White+bad and Black+good concepts.

To set the stage for the main discussion, I will start with eight results. There should be no disagreement about what the data are, if we share assumptions about the nature of the test, and in fact these results are now sufficiently commonplace so as to be accepted. The real discussion we might have lies in the material that follows, concerning some unexpected effects and what they may be telling us about the nature of implicit attitudes. As well, I will mention a couple of the larger set of questions that remain unanswered.[7]

Implicit Race Bias Using the IAT: The Main Results

1. Many White Americans and Asian Americans in the world show robust association of white+good and black+bad compared to the opposite pairings; the percentage of White Americans who show such an effect is upwards of 75% with the remaining 20-25% being neutral or showing slight Black preference.

2. On self-report measures of group preference, White Americans report positive attitudes toward their own group but far less so than observed on the IAT.
3. African Americans and Africans do *not* show matched and opposite positive associations with their own group (i.e., Black+good). About 40% show pro-white bias, a little less show a pro-black bias, and the remaining 20% show neutrality.
4. On self-report measures of group preference, Black Americans report positive attitudes toward their own group and far more strongly than that observed on the IAT.
5. Self-reported and implicit attitudes towards the two groups are correlated, i.e., those with stronger expressed anti-black sentiment also show weaker associations of Black+good relative to White+good on the IAT. This correlation is consistently obtained, is moderate in size, suggesting that self-report and IAT measures have some shared components and have unique aspects to in what they each measure about group preference.
6. To respond to questions about whether the obtained results are a function of the laboratory environment in which they are experienced or the self selection of those who arrive at the website, a database continues to be gathered of published and unpublished studies that use "real world" sample populations (of actual voters in elections, doctors prescribing medication, nurses who work with drug addicts, managers who are making real job selections between White Europeans and Arabs, and so forth). The growing database of 56 studies at present includes a subset of studies of attitudes and stereotypes of racial and ethnic groups.[8]
7. Performance on the race IAT predicts behavior toward members of the groups tested. In a meta-analysis, the IAT race bias predicted such behaviors as non-verbal expressions, ratings of ability and performance, and decisions about resources. In each case the greater the bias detected on the test toward a particular group (Jewish Americans, Black Americans), the greater the negative attitudes observed in behavior toward them. The studies included in the published meta-analysis are largely samples of convenience, of college students who are more homogeneous than the population at large and the reported correlations may be an underestimate of the actual correlation between IAT score and behavior.[9]
8. Research shows that the IAT race effect is malleable, although the extent of malleability and the precise conditions under which it is obtained remain open questions. Malleability of any sort was not initially predicted because performance on the IAT is difficult to adjust based on one's volition. It takes much practice and an understanding of the task to move the outcome in the direction of one's preference and even so, it is not always possible to be successful. That is the nature of the test, and its signature feature. However, studies have shown that particular interventions can produce a significant reduction in the baseline IAT race bias. Such a possibility tells us that even those group attitudes, like race attitudes, that have built up over extended periods of time and been in us for extended period of time (unlike attitudes toward a new movie) are pliable in the direction of the intervention.

Unexpected Results

A student once described to me his frustration at the way research was being done in his lab. He felt that his advisor conducted a large number of experiments most of which were a wash. When a study did work, they published it. The student had a different view of how science should be done and wondered whether I considered this to be tricky or problematic science practice. Knowing the research program well, my response was more reassuring than he may have expected. The advisor had a strong theory, but a weak method, I thought. When tests didn't support the theory, it was reasonable to distrust the failure and continue to believe in what is a gorgeous theory. I offered my own research program as and example of the opposite problem. I had less of a theoretical stake in the questions I asked (okay, I'll admit it, I don't actually have a theory; I sure hope though that the work is theoretical, in the sense that a particular set of problems with particular histories and assumptions have been selected for study). What I did have was a preference for a classy family of methods in the form of priming, the IAT, startle responses, neuroimaging, ERP and the like. I have banged each can with one of these, and then whatever data were shaken out needed to be dealt with. The situation was more akin to looking at a visual illusion. We wouldn't say the illusion is wrong or that it didn't work like it was supposed to. It is what it is. It needs explaining.

Still, it is natural to go into most experiments with expectations of what ought to be, and every now and then we are puzzled by what comes out. This then, is the stuff we may find worth discussing.

Developmental Evidence

The work that has surprised me most recently came from research I began soon after arrival at Harvard in 2001. Here I met two amazing developmental psychologists who had also just arrived. I fell in love with Liz Spelke and Susan Carey, especially their approach to understanding the mind. I realized that by contrast to what they knew about their phenomena, I had no clue as to where implicit attitudes, beliefs, or identity, originated or how they developed. By contrast, I had never had a subject who needed a mother's permission to be in a study, except the occasional 17-year old summer student. But that was preposterous. How could I have gone for this long without worrying about the origins of implicit social cognition? Starting about a decade ago, Andy Baron, then an RA and later a graduate student in the lab, and now a professor at the University of British Columbia made a discovery that I found to be surprising. It was followed by confirmation in research by another student, Yarrow Dunham.

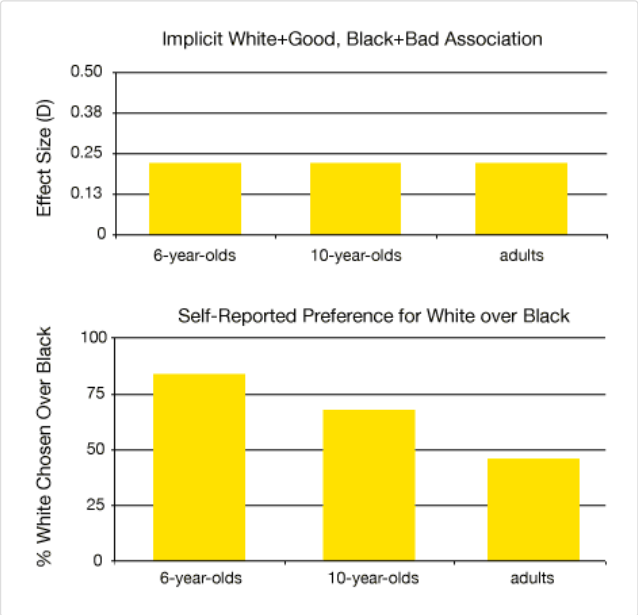
An infant is born into the world with some obvious built-in preferences — it spits out bitter tasting food and sucks in the sweet, it knows to orient towards a touch applied to the cheek and it imitates the facial expressions of other before it. We also know that babies have a ready mechanism to learn new preferences and that one dimension along which learning proceeds is familiarity; that which is seen over and over again, and by extension, that which is akin to the familiar is preferred because it, by comparison to the strange and exotic, is assumed to be trustworthy. However rapidly social learning occurs, it is assumed that learning requires some experience and looking at the same concept, say an attitude, develop over time provides a rich source of learning about its nature. Nobody would say that a three year old and an adult have had the same experiences with race. Even those who live in the most shielded of social environments in America have got to "learn race" in some way — by observation, via stories, through education, and perhaps most persistently from the media. It goes without saying that American adults have many hundreds of thousands of extra units of experience with social groups in general, and on race in particular than young children. We went in without strong hunches but some expectation that we would see developmental shifts. The most likely effect we would see is that as experience in intergroup settings and knowledge increased with age, so would implicit race bias (stronger pro-White/anti-Black associations).

As the graph at right shows, that was not what we obtained.[10] When asking the blunt question of whom one likes, the youngest White children we tested, aged 6, are significantly likely to say that they prefer a White child to a Black child. This fits with research by Francis Aboud, who showed much earlier that young children are not pure and without explicit prejudice.[11] Interestingly, 10 year olds in our sample are less likely to show as strong ingroup preference, suggesting that they have either lost that strong preference or more likely, that they have learned about the inappropriateness of expressing consistent preference for those from one's own group. We have some anecdotal evidence that 10 year olds may indeed be struggling here — an occasional child did ask the experimenter to remind them about the race of the child on the previous trial so that they could choose one from the other group! As the data show, adults are at chance in their choice, with Black and White targets selected equally often. Together, this progression towards conscious egalitarianism may be best viewed as an effect of social learning about race tolerance.

The IAT data were unexpected. Not so much that we see evidence of ingroup preference at the earliest age we were able to test but rather in the *stark stability of the preference across development*. Year upon year of greater experience with race seems to have no impact on the IAT attitude. We are left with the conclusion that whatever the nature of the preference detected by the IAT, it is a primitive form of attitude that expresses itself relatively early in life and doesn't change with increased experience with the attitude object that must occur between age 6 and adulthood.

In further work with Yarrow Dunham, we went to even younger ages, something we could not achieve with the version of the child IAT that Andy had developed. Now, we adapted a

test previously used by Hugenberg and Bodenhausen in which racially ambiguous faces were judged as Black or White. Results with adults had shown that if the racially ambiguous face happened to have a smiling expression it was more likely to be categorized as White, and when the face had a frowning expression, it was more likely to be categorized as Black. Yarrow thought this would be a good procedure for our purposes because even a child can do it.



So we conducted a large n study including children as young as 3 years of age, ramping up to adults and including a range of age groups in between. We found again, in this new test that is quite different from the IAT, the same striking result: White children as young as three and all ages upwards through adulthood showed an identical level of ingroup favoring bias.

Questions: Why is this effect invariant across the age spectrum? Shouldn't the age of the person matter? If age doesn't matter, it seems obvious that experience doesn't shape this particular form of attitude. It gets into us early and stays. I welcome views on the meaning of this result and whether other procedures can be used to go to ages even younger while still being feasible for use across all ages for comparison.[12]

A second result has been on my mind since we first obtained it at Yale in the mid 1990s. When we sliced the data from 50 Black American students we found that the no-bias absolute zero sat squarely in the middle of the distribution — that is, half the sample of Black students fell to the right of the zero score showing a pro-white bias, the other half to the left of the zero, showing pro-Black bias.

The first such observation of a lack of strong ingroup preference led us to interpret these data as unique to the sample. Black students at an elite institution with a largely White faculty, administration and student body, we speculated, may show such a pattern of neutrality, but surely it wouldn't obtain in most other samples of Black Americans. But the very same result has indeed obtained over and over again and it has been replicated among Black South Africans as well (the study was motivated by the recognition that Black Americans may show this because they are not only disadvantaged but also because they are a statistical minority in the United States, something Black South Africans are not).

At the website, with hundreds of thousands of Black Americans taking the test, about 40% percent of Black Americans show a pro-White effect, about 40% show a pro-Black effect, and the remainder are neutral. Each ordinary sized sample we test, as well as the large numbers at the website show the same pattern. Here's the whammy: Young Americans who come from disadvantaged groups look like the adults of their group.[13] Again, we might have expected that children would show an ingroup favoring effect that is then mitigated in adults who come to realize that their group isn't regarded to be as good as others. But younger children who are Black and Hispanic show effects that map onto those produced by adult members of their group. How is this possible? How do children come to know and internalize so early in life exactly what adults have? Shouldn't there be an effect of slower learning that is visible in developmental trends?

These data suggest an additional possibility that the data from White samples do not. At least the data from White children and adults can be explained as reflecting an early emerging and stable preference for one's ingroup, with that preference reflecting a socially adaptive mechanism. When adults of a group do not show the expected group love, we assume that they have come to terms, through experience, with lower regard for their own, less privileged, group. But children of that group ought to show robust ingroup preference, before it gets beaten out of them. We don't see that; instead, we see a fateful mapping of neutrality between child and adult. Implicit attitudes represent an automatic and early internalization of whatever the social world signals about the status or value of one's groups. That's the reason Black and Hispanic young people show exactly the same reactions as their adult group members.

A final age-based result gives further food for thought. Among the many tests we've used at the website, a long-standing test is one that measures the stereotype that men belong in the world of work, women at home. The test involves association of clear gender markers like first names to capture the categories female (Susan, Mary, Jane) and male (Steve, Matt, John) with attributes of home (garden, kitchen, children) versus career (job, briefcase, office). In adults, both male and female, the outcome is a robust association of male+career and female+home relative to the opposite pairing. But there is also a clear age trend on this test. Starting with children in the low teens, the data from thousands of subjects show that as age increases, so does this stereotype. Not surprisingly, the story we and others tell is that the world experienced by younger people is indeed different from that of older people on this dimension of gender and career — there are many more women in the workplace with each passing decade; even the number of stay-at-home dads is likely to be higher now than before. This difference in environments, we have assumed, is responsible for the observed age effect.

In the same breath, many parents report that likewise, on race, their children are in a quite different world than their own. They report stories of their children being oblivious to race, having friends who cut across race lines, and rap and hip-hop having changed their children's attitudes toward Black Americans. We don't like busting parents' myths, but the data do not bear this out. Unlike the gender stereotype test, the data from hundreds of thousands of test takers at the website shows no difference in the race attitude as a function of age. Whatever social changes have occurred that involve race, our children are not different from us in their implicit race attitude. What does this mean, given the change in gender stereotypes by age? Does it mean that in spite of all the changes since civil rights legislation, social change and media change, that a 10 year old and a 70 year old have the same race attitude? Does it mean that racially our lives are still so segregated that that to nudge implicit attitudes we haven't created the appropriate conditions of contact?

Self-Esteem

Many years ago, Jennifer Crocker and Brenda Major reported a result that was unexpected. Across subgroups that varied greatly in social advantage, they found that members of less advantaged groups did not show lower self-esteem; in some cases, it was actually higher.[14] The review they presented was thorough and left little room for argument, except one: their data, due to availability, consisted of self-report measures of self-esteem asking for reactions to questions such as "I am a person of equal worth compared to others".

It is possible that on such blatant measures, members of socially disadvantaged groups are especially likely to take control by expressing a positive sense of self. I certainly thought that was a possibility and when we began research with the IAT and an implicit self-esteem measure was developed (strength of me+good vs. me+bad, compared to "other") we looked at self-esteem by racial groups and found the same surprising result (at least to me) that Black Americans showed the highest self-esteem of any group. Another intuitive prediction gone by the wayside! Self-esteem is robust enough in most people that "given" group membership, like race, does not seem to affect it.

Many years later, we showed another twist to this result. A great deal of research had shown that on self-report self-esteem inventories, Asian Americans and Asians reported lower regard for self, and many cultural psychologists believed this difference to be reflecting a serious western versus eastern difference in self regard, i.e., not a simple effect reflecting demand characteristics to show confidence or modesty but a true difference in how the self is viewed. Our work in China, Japan and the United States showed that whatever the differences in self-reported self-esteem were, people of all three cultures showed robust implicit self-esteem. An unexpected result based on intuitions of the past but giving a more accurate view that reports that are not modified by the pressures of presentation show that a tendency toward high self-esteem is universal. Here, the implicit measure wiped out a previously established cultural difference and showed that we are all more similar than may have been assumed.

We Don't Know Why Yet

In reporting on some well-replicated results earlier, I showed that even on tests on which a large majority within a sample show a particular preference or belief (White=good, Home=Female) there is always a significant minority that shows no such preference or a slight tilt in the opposite direction (Black=good, Home=Male). So who are these people who on a given test do not show the effect that the majority of the sample (70, 80, 90%) do? Why, if they apparently live in the same culture as the rest of us, do they not show the same preferences and beliefs? In our conversations we find that those who do not show the effect demonstrated by the majority also say that they do not have any idea as to why they are unique. It is my sense that those who are neutral are cognitively oblivious to signals of social group membership, either on a particular dimension such as gender or race/ethnicity or more generally.

A White construction worker who was once interviewed regarding his IAT result (he had showed no bias one way or another) said that he didn't notice things like that. His spouse was surprised that in identifying a co-worker, he hadn't immediately identified the co-worker as Hispanic, which may have been the most obvious way to single him out from the group. Our construction worker said that he didn't notice things like that. It is possible that some of those who do not show the majority pattern are oblivious in this way. Others tell more expected stories about trying to lead an examined life around social issues and practicing counterstereotypic behavior. But we have no way of knowing whether this actually accounts for their IAT behavior. So far, we have found no clear story we can tell about the qualities of those who don't show the standard effect of ingroup preference (we haven't had the time to do intensive studies of this), it is an oft-asked question to which I don't have a good answer.

A second dimension that I put forward for discussion is the question of malleability. I know that I began as a skeptic on this, believing that intense and lengthy interventions would be needed to shift implicit cognition. To my surprise, my colleagues proved that even milk toast experiences that lasted a few minutes were sufficient to shift implicit attitudes. Such effects have been found since then in my lab. The presence of a Muslim experimenter made attitudes toward Arab Muslims more positive in liberals and more negative in conservatives. Thinking about the virtues of winter for a few minutes blunted an otherwise positive attitude toward summer. But I don't know how easily replicable these effects are and I certainly don't know how to definitively produce an effect or a long-lasting one. Maybe others do and they'll tell me in their comments[15]. I wonder when we'll have a clear understanding of the conditions under which we can definitively observe and predict change, and how we might mold it into a state desired by our conscious attitudes.

In the age of Obama, it may be passé to quote MLK Jr. But until the president develops a spine, I'll stick with the older civil rights leader. In a speech that Martin Luther King Jr. gave as among his last (at the convention of the American Psychological Association) he said that what social scientists could do to assist the cause is to "tell it like it is". We have told it like it is about the research discoveries, whether it is palatable or not. The personal discovery I made about the contents of my own mind, especially regarding race, hasn't been easy to come to terms with even though a long time has passed since I first came face-to-face with it more than 15 years ago. But there were others, colleagues and strangers who also told it like it is about themselves, and for their company I am grateful. They have made it no longer a matter of courage for me to tell it like it is.

Notes

1. See, Warrington, E. K., Weiskrantz, L. (1982) Amnesia: A disconnection syndrome? *Neuropsychologia*, 20, 233-248.
2. Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227-234.
3. Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106, 226-254.
4. Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 5-18.
5. von Hippel, W. and Trivers, R. (2011). The evolution and psychology of self-deception. *Brain and Behavioral Sciences*, 34, 1-56.
6. Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.
7. Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18, 36-88.
8. **http://faculty.washington.edu/agg/** (follow link titled "Studies showing use of the IAT with "real world" subject populations")
9. Greenwald, A. G., Poehlman, T., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97, 17-41.
10. Baron, A., & Banaji, M. R. (2006). The Development of Implicit Attitudes. *Psychological Science*, 17, 53-58. See also, Rutland, A., Cameron, L., Milne, A., & McGeorge, P. (2005). Social norms and self-presentation: Children's implicit and explicit intergroup attitudes. *Child Development*, 76(2), 451-466.
11. Aboud, F. E. (1980). A test of ethnocentrism with young children. *Canadian Journal of Behavioural Science*, 12, 195-209.
12. Of course, we can conduct yet other tests with even younger children who are unable to understand and respond with language. In infants preferences cannot obviously be measured with the procedures used here. Instead they can be measured through reaching, grasping, and looking time. But such measures cannot be meaningfully applied across the life span. And to speak to the question of attitude stability and change over the course of development, we do need the same measure to be applied at all age levels to learn anything of interest about stability versus change.
13. Dunham, Y., Baron, A., & Banaji, M. R. (2007). Children and social groups: A developmental analysis of implicit consistency in Hispanic Americans. *Self & Identity*, 6, 238-255; Dunham, Y., Baron, A. & Banaji, M. R. (2008). The development of social cognition. *Trends in Cognitive Science*, 12, 248-253; Newheiser, A. & Olson, K. R. (in press). White and Black American children's Implicit Intergroup bias, *Journal of Experimental Social Psychology*.
14. Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review*, 96, 608-630.
15. And an ongoing experiment at University of Virginia, issued as a challenge to design any five minute intervention that can change the race IAT, is producing some interesting data that will soon become public on the types of interventions that work and do not.

Acknowledgments: I thank the Edmond J. Safra Center for Ethics at Harvard University for support and Paul Meinshausen for comments and editorial assistance.

September 19th, 2011 | Categories: Humans, Participants | Tags: race, science, sociology |

9 comments to The Dark Dark Side of the Mind

David Hingston
September 20, 2011 at 4:58 pm

Could it be that you're not talking about race so much as color? From day one, I should think, our experiences of light vs dark (day vs night, sunlight vs shadow, open meadow vs cave) associate white with what tends to be pleasant, and black with what tends to be unpleasant.

For what its worth, I am a 61-yr old white male. I grew up in a Lilly-white community — Palo Alto and Los Altos, CA. I couldn't have been any older than how old I was in the second grade but might have been much younger when I noticed my first black person (a man). He and I were both standing in the aisle of a grocery store. Apart from the fact

that he was much taller than I was, what struck me most what how white his teeth were. I was envious.

Susanna Siegel
September 25, 2011 at 11:13 am

Thanks for the fantastic article. I'm wondering what you take to be the main theoretical options for what kind of implicit attitude the IAT reveals. What people actually do on the IAT is associate one concept with another. But we can still ask what kind of underlying psychological state explains that behavior. In my mind this raises two main questions.

Q1. Is there evidence for whether underlying state is:

(a) an association of concepts, on par with "hot-cold", "salt-pepper", "abbot-costello", "smoke-fire" — movements of the mind from one concept to the next that lack the predicative structure that would make them the kind of state that could be true or false?

…or whether it the underlying state includes:

(b) an attitude that attributes properties to the groups in the test? On this option, someone who shows strong association of white+good and black+bad does so because they have an implicit attitude attributing negative properties to blacks and positive ones to whites.

What kinds of experiments do you think would speak to this question?

Q2. Another dimension of the underlying attitude is negative affect. What do you think the main theoretical options are for where negative affect fits in the implicit attitude?

It seems to me that it could be either downstream of a truth-evaluable attitude (eg, you have negative affect toward a group, because of what you covertly believe about them), or it could be upstream of it (you covertly believe they have negative features, because of your negative affect).

Perhaps something like these same downstream/upstream options apply in the case of association: either implicit bias starts life as negative affect, and grows into an association from there; or the associations come first, and negative affect comes after.

Is there developmental evidence that speaks to whether negative affect toward a group is developmentally prior to full-blown implicit attitudes?

Thanks again
–Susanna

Eric Mandelbaum
September 25, 2011 at 5:53 pm

Many thanks for the article. A few questions come to mind (especially the propositional vs. associative structure question, but Susanna seems to have that covered):

1) Have you (or any of your colleagues) tried correlating individual differences in the salience of race (as a perceiver) with IAT scores? You note that anecdotally it appears that people who somehow avoid having an implicit bias also seem oblivious to race in certain ways. This seems like a testable hypothesis. For example what if you gave a description task where you introduce subjects to characters of different races, professions, and backgrounds, and ask the subjects to describe those characters after some waiting period. If those who mention race later on in their description (or don't mention race at all) correlate highly with a neutral IAT preference, that seems like it may be a start to showing the anecdotal hypothesis holds. Do you know of any studies that have attempted running something like that?

2) What do you think the relation is between implicit bias and weapon bias that Payne finds? Payne cites that the correlation between those who show a high implicit racial bias and those who fall prey to the weapon bias is low, which I found quite surprising. Do you think that the two effects are tapping into different cognitive systems or attitudes? Does the weapon bias show a different developmental curve (or lack thereof) than the one your report for the IAT?

Thanks again for the great read.

Louise Antony
September 28, 2011 at 2:04 pm

I'd like to add a couple of notes to Susanna's and Eric's comments.

First, if it turns out that the IAT is tracking conceptual associations, rather than revealing propositional attitudes, that might explain some of the puzzling features about the IAT findings that you mention, like the stability of IA's through development, the matching of results between children and adults, and the relative ease with which the associations can be altered. Those things might be expected if the IAT is revealing what the subject knows about the content of a stereotype, rather than beliefs or values the subject holds.

Second, insofar as the issue above is unsettled, I'd like to urge much more caution in characterizing IAT results in terms of "implicit bias," especially in cases where those results appear to contradict the subject's avowed values and attitudes. It cannot be just assumed that the IAT results constitute evidence against the subject's own reports — in advance of independent evidence, the mismatch could equally well be taken as evidence that the IAT is not tapping into stable or deeply held attitudes. I say this not because I think that we human beings are infallible about our own mental states — demonstrably, there are cases in which our own motivations are opaque to us. But we also know that associations can outlive changes in beliefs and preferences. (Having been an atheist now for nearly forty years, I find I still think "full of grace" to myself whenever I'm introduced to someone named "Mary.") In cases where a person has devoted a great deal of thought to an issue, or when a person has made a great effort to change her behavior, there is reason to think she knows her own mind, and we should not be casual about contradicting her. The charge of "bias" is incendiary, even if it's explained that the bias is presumed to be unconscious, and hence not something for which the subject is necessarily responsible.

So like Susanna and Eric, I'd be very interested to hear more about the way you are thinking about what the IAT measures, and I'd welcome more information about studies that examine the relations among IAT scores and behavior that would independently support a charge of bias. Eric reports finding Payne's results surprising, but they seem to me to be rather what one would expect if the IAT is not really tapping into the cognitive and affective states that control and explain much of our behavior.

Thanks for the article, and the opportunity to comment!

aa.second
September 28, 2011 at 2:53 pm

I'm using an assumed name because I really don't want this to see self-congratulatory in any way. The email address is real, however. I do question the idea that those who race neutral on the IAT for race are not race conscious.

I test neurtal on the IAT for race, and yet I think I am very race conscious. Race is an annonying salient feature in my experience, and I make some effort to keep this somewhat under wraps.

Here's what I think produced the neutral reading. I grew up in a white community in a very racist part of the country. In approximately 1986, sitting in our car at a stop light on Nassau Street in Princeton, my adolescent son asked me if I have noticed how much better looking blacks are than whites. In general, of course. I saw immediately what he meant and that it could easily be true. And that the same idea would never have occurred to me. I felt so ashamed that my perception was so skewed and racist. So I started working on it. I tried to uncover the narratives in my mind that would go on when I'd see/meet/read about a black person. I worked quite hard to understand some of the factors behind features of some blacks that might draw criticism and substitute explanations that did not denigrate them. And I got active in trying to promote diversity in my university (not Princeton).

That worked. I was surprised at the IAT result, especially since I clanged on women and science, about which I care a great deal, but not VERY surprised. But then I haven't done the work on women and science. E.g., I've sat on hiring committees and simply noted the prejudices about women and science I have without really trying to alter them. (Not voting on them is hard enough!)

Arthur G. Miller
October 3, 2011 at 9:45 am

THE DARK DARK SIDE OF THE MIND

Banaji's article, in many respects, is exactly what we would expect from an extremely prominent social scientist writing about something she really knows about—namely, a methodology (the IAT) which she helped discover, and which has generated a voluminous data base to which she has been a continuing contributor. The article is superbly written, personally engaging, compelling and forceful in its insights and contextual perspectives, and simply loaded with interesting information—major findings, conceptualizations, and several provocative unanswered questions waiting further inquiry. Indeed, this inquiry will doubtless occur, as I find the essay so stimulating that I can easily see it having a research-stimulating effect—perhaps the highest praise one can bestow upon a review of research. As for her substantive observations, there are many to consider. For myself, the stability of the IAT for race (but not gender) across a vast age range—what Banaji terms the "stark stability of the preference across development– is as interesting a set of data as one is likely to encounter, and the group differences between white and black populations beg for an explanation, as do the individual differences in IAT scores about which Banaji ponders at some length. Individual differences are particularly important not to ignore in a paper that is devoted primarily to strong group effects and, it would seem, generalizations about the mind itself.

Allow me, however, to throw in here a couple of tiny caveats. First, I think Banaji could have left out the dig at President Obama, i.e., her phrase in the last paragraph, "But until the president develops a spine…" I'm hardly defending the president here, nor necessarily refuting Banaji's observations regarding his orthopedic shortcomings. Nevertheless, I think the phrase does not add to the force of her paper which is both very readable and sophisticated, and it is also a bit distracting, as I found myself musing about her motives or reference point, i.e., just what is it about the president or his policies that Banaji finds so onerous. And once one is prompted to play this guessing game and think about such things, one is not thinking instead about the many other things that Banaji is so uniquely qualified to tell us about. Perhaps I am over-reacting here—the phrase occupies 7 words.

A second issue is perhaps less of a criticism than an elaboration of a point well made. Banaji frames her paper, at least in part, by commenting upon the courage it took for Patricia Devine, as well as Banaji and her colleagues, to engage in the probing of unconscious racism in white people, what she terms "the hidden biases of good people." She does this very early in the paper, as well as at the end, where, after quoting from Martin Luther King Jr., she concludes, "We have told it like it is about the research discoveries, whether palatable or not." Having taught many undergraduate courses on the subjects of stereotyping, stigma, and prejudice, I agree wholeheartedly with Banaji's point. It is not easy to tell "good" people that they are not quite as good as they could be on matters pertaining to race (and related social/ethnic/religious groups). Many people do not like this message and they will not like the messenger either. However, I think it is worth noting here that many social psychologists have told it "like it is," have discovered unpalatable truths, and continue to do so. Historically, the names Stanley Milgram and Philip Zimbardo come immediately to mind, but many others could be added to this large list. Banaji is clearly not claiming a courageous or heroic status solely for the research at issue in her paper, but I do think it should be recognized that many social psychologists have told it like it is, and often paid a severe price for doing so in terms of criticisms leveled at them not only from resistant lay persons but professional colleagues as well. (The same would be true undoubtedly for many research projects in the natural sciences). To a degree, of course, this is precisely as it should be. Some might say the angst instigated by many scientific discoveries is simply part of the scientific enterprise, a natural (if at times vexing) reaction to unexpected, emotion-laden findings.

Finally, I would like to comment on the title of Banaji's paper. Titles are interesting phenomena in their own right. All authors want their work to be read, and titles are undoubtedly chosen in an effort to attract readers. Banaji's title, in my view, is particularly interesting. In fact, when I first read the title, I thought it read "the dark side of the mind." It took a second reading of the paper before I caught my mistake—I had missed the second "dark." The title is a kind of double entendre. In one sense, the title's meaning is clear and straightforward. Referring to her colleague Patricia Devine's classic study on automatic and controlled processes in racial prejudice, Banaji describes it as a "breakthrough that would open a brand new gate to theorizing and experimentation about the double dark side of the mind." One of these "darks" refers to the unconscious, and the second to Blacks ("humans with high amounts of melanin in their epidermis"). Thus, taken literally, Banaji's title is simply a somewhat imaginative description of the unconscious racial bias that is an essential focus of the paper.

However, I find another aspect to the title, beyond the literal one noted above. Banaji seems to be addressing human nature itself. That is, she is indicating that we all have a dark (even dark dark) side to our minds. Most of us are, of course, motivated to think rather highly of ourselves, and social psychologists seem to have a penchant for describing most people as good. Consider, for example, Banaji's observations: "Psychologists and neuroscientists have found it profitable to think about the mind's fractures by relying less on the malign motives of bad social actors as the explanation, and more on elucidating the evolutionary presses that created the minds we have and the sociocultural and situational presses that exert influence on us more proximally. Over the past 25 years, we have studied the hidden biases of good people, i.e. self-professed egalitarians."

After considering the research in this paper (and the innumerable other instances of good people behaving badly in social psychology experiments for the past 7 decades or so), I am left to wonder a bit about this idea of "good" people." Who exactly are these people? I guess it depends on the meaning of "good." I think we can agree that Banaji's focus is not on "the malign motives of bad social actors," presuming some agreement on her precise meaning of "malign" or "bad." To the degree that research participants are not aware of their actions in the IAT laboratories around the world, there would certainly seem a correspondingly low degree of intentionality. But, that aside, there clearly is plenty of bad behavior to be seen here. Banaji suggests that her participants are largely good social actors without malign motives. Are we to ice the cake further and let them off the responsibility hook as well? I think a quick answer here is "yes," a somewhat unsettling note upon which to end this commentary.

Arthur G. Miller

Jenny Saul
October 3, 2011 at 10:35 am

Hi Louise,

Anthony Greenwald has a downloadable pdf of studies showing connections between IAT scores and real-world behaviour. You can get it here: **http://faculty.washington.edu/agg/**.

Andy Baron
October 4, 2011 at 11:33 am

This is a terrific article both broad in scope and specific in findings. Several decades later we know much more about the nature of implicit social cognition. The set of results detailing the early emergence of these implicit attitudes and their stability across development is quite striking. A few questions come to mind.

What advice would you give parents and educators concerned with raising more tolerant beings?

What about the first 3-6 years of life primarily contributes to the formation of these attitudes?

Do you think that implicit attitudes toward non-social/human objects (eg. food, animals, etc) would be similarly stable across development? Or, is the stubbornness to change something particular to how our mind treats other people?

Patricia Devine
October 5, 2011 at 9:06 pm

Coming out of the Darkness

What a wonderfully thought provoking essay and impressive body of work! I remember well my early conversations with Mahzarin about my interest in examining the effects of unconsciously priming stereotypic concepts of Black Americans. My friend did not give away her surprise or concern that I might be exploring something quite controversial let alone illicit or dangerous. I certainly had no sense of the potential perils that provoked her trepidation. My goal was to understand why those who renounced prejudice – self-professed egalitarians – showed evidence of bias that belied their conscious intent. It seemed to me that the main challenge was to understand the dissociation between what people consciously believed about their attitudes and what their behavior showed (Crosby et al., 1980). Other theorists championed the theme of the inevitability of prejudice, a theme that seemed tremendously pessimistic and antithetical to the idea that people can change. Could it be possible that there was something below the surface that was driving the inconsistency between conscious and unconscious responses? To explore processes below the surface required asking questions about and using methods to detect these unconscious processes.

What I remember most about my discussions with Mahzarin was that she was interested in what I was doing and was supportive. (And, she clearly understood my need to do research on impression formation!) She asked deep and probing questions and quietly encouraged me to explore this as yet unchartered territory. Her support was reassuring because neither my mentor nor other faculty members at OSU at the time seemed the least bit interested in what I was doing. What I didn't know was that these early conversations and my early work would encourage my dear friend to join the effort to explore and understand what she refers to as the "dark side of the mind."

Here we are all these years later, both still interested in "trying to figure it all out." The Implicit Association Test (IAT) has been a tremendously useful tool in directly assessing the presence of implicit biases. Prior to its development, much of the reasoning about the impact of implicit biases was based on indirect evidence that was consistent with the idea that unconscious processes affected perception and behavior. Being able to directly measure these biases has enabled researchers to establish the prevalence of these biases and brought into focus puzzling patterns of data that Mahzarin so eloquently reviewed and challenged us to think about. Having limited time to respond, let me comment on just two issues that have puzzled Mahzarin as she has explored the nature and consequences of implicit biases – the developmental stability of implicit race preferences revealed by the IAT and the question of such biases being malleable. The issues are, I think, very much interrelated.

Mahzarin and others have been struck by the stability of implicit race preferences among Whites across development. That is, whereas explicit responses fall more in line with socially prescribed norms over time, with the majority of people expressing increasingly egalitarian views as they move from childhood to adulthood, implicit biased favoring Whites over Blacks remains high and constant over development. I don't find this finding surprising or puzzling. We have known for a long time that Whites show bias favoring their group from a very early age. The IAT and other methods have simply revealed to us that the biases that were originally assessed with self-report measures can also be revealed with implicit indicators. Whatever learning processes lead to the development of implicit preferences for Whites over Blacks, the primitive attitude to which Mahzarin refers, produce strong effects on the responses to which we have conscious access and those to which we do not. Social leaning about race tolerance and egalitarianism leads people to examine their beliefs and to consciously wrestle with whether the racial attitudes align with more abstract principles of equality. For those who sincerely embrace egalitarian values, their conscious beliefs change. The problem is that they are not likely to be aware of their unconscious processes or the need to made adjustments that would prevent them from showing bias in unintended ways (i.e., they may not understand the need for work to undo early learning about what the social world signals about the status or value of social groups). Consequently, a great many people are unwittingly complicit in the perpetuation of discriminatory outcomes. People simply won't (can't) change something they are not aware of!

This perspective, however, raises the question of whether people would be motivated to make efforts to rid themselves of unconscious biases if they were made aware of them. And, if motivated, what should they do. For many years, research in my lab showed that self-professed egalitarians struggled with knowledge that they don't always live up to their egalitarian standards. When confronted with evidence of their bias, they felt guilty. The guilt motivates them to expend effort to understand why they show bias and to be interested in information about reducing bias. This evidence is compelling but indirect. They seemed sincere but would they work at overcoming bias? Yes, they will. When their bias is revealed to them using the IAT, those who are personally motivated to overcome prejudice will spend time on a task they are told will help to eliminate unconscious prejudice. Again, the evidence, though compelling, is indirect – we provided no evidence that implicit prejudice was reduced. More recently, we developed and tested the effectiveness of an intervention designed to (1) increase awareness of implicit bias (2) provide education about the nature of implicit bias (likening implicit to habitual responses) and (3) teach strategies that if practiced would help people reduce the prejudice habit. The goal was to help participants become sensitive to their own implicit biases and then equip them tools to combat the biases.

To raise awareness of their implicit bias, our participants completed the IAT and received feedback about their performance. As is true in the literature, the vast majority of our participants showed implicit bias favoring Whites over Blacks. Control participants were then dismissed. Those in the intervention group watched a 45-minute narrated slide show that included the educational (i.e., explanation of the IAT, summary of evidence linking the IAT to discriminatory outcomes in a wide range of contexts) and training (i.e., stereotype replacement, counterstereotype imagery) components. The training strategies had previously shown to produce incidental reductions in implicit bias (i.e., the strategies were practiced at the behest of the experimenter without any intention to reduce bias or awareness that the strategies may produce lead to reductions in implicit bias). Could these techniques be used proactively to bring one's implicit responses in line with their conscious intentions? We brought participants back to the lab on two additional occasions, 4 and 8 weeks after the initial IAT assessment, during which we administered the IAT (without feedback).

The intervention was successful! At the outset, the level of implicit racial bias was equivalent for our intervention and control groups. The level of bias remained unchanged for the control group participants – despite having been made aware of the implicit bias. However, by week 4 and extending through week 8, those in the training group showed a dramatic drop in their IAT scores. This study provides the first evidence that people can use the power of their conscious minds to reduce implicit biases. It seems clear, however, that awareness of one's implicit bias is not sufficient to break the prejudice habit, though it may be necessary. It appears that people also need guidance on how to break the prejudice habit. Once equipped with strategies, those concerned about discrimination and their role in perpetuating it, however unwittingly, will do the work to reduce

their biases.

So, is the consistency in the magnitude of implicit bias unusual? I would say only to the extent that one assumes that changes in beliefs translate immediately to unconscious processes. And there is good reason to believe this assumption is not valid. Are implicit biases malleable? Previous work has shown that they are but had left unknown whether (1) the reductions in bias were enduring or (2) whether people could use strategies intentionally with the goal of reducing implicit bias. For those who are personally concerned about ridding themselves of bias, the news is good. A little hard work directed intentionally toward the goal of reducing implicit bias can bring unconscious responses in line with conscious intentions and shine light into the darkness.

(posted for Patricia Devine by G. Comstock)

« **Animal In Mind: People, Cattle and Shared Nature on the African Savannah** **The Health Impact Fund: a better way to reward new medicines** »