Mahzarin Banaji and the Implicit Revolution

January 31, 2018

The concept of unconscious thought — that there are aspects of our minds that we are unaware of but that nevertheless influence our behavior — has been around since the days of Descartes, but only in the last 30 years have psychological scientists put these implicit cognitions under the proverbial microscope to examine where they come from, how they work, and how they relate to perception, learning, memory, judgment, and behavior.

On the frontlines of this revolution is APS Past President Mahzarin R. Banaji, the Robert Clarke Cabot Chair of the Department of Psychology at Harvard University. A panel of Banaji's collaborators and former students, esteemed psychological scientists in their own rights, gathered to discuss the influential role Banaji has played in their research at the 2017 APS Annual Convention in Boston. The symposium was in honor of Banaji receiving the APS William James Fellow Award.

The Mind's Projectionist

APS William James Fellow Anthony Greenwald, professor of psychology at the University of Washington, became Banaji's advisor when she began her graduate studies at Ohio State University in

the 1980s and remains her most prolific collaborator to this day. Their theoretical and empirical work led to the 1998 creation of the Implicit Association Test (IAT), a pioneering assessment tool that has changed how we understand and measure unconscious attitudes. The two also have contributed to countless other scientific breakthroughs and published myriad journal articles.

While the idea of two separate mental levels — a higher, conscious level and a lower, unconscious level — was far from original by the late 20th century, the pair's specific characterization of the relationship between these two levels was indeed revolutionary, Greenwald said. He described the connection thus:

"The conscious level of the mind is obliged to use what the automatic level provides to it, and in this way the lower level controls conscious perception, thought, and judgment." Greenwald likened this relationship to that of a theater projectionist and a film audience, where the former exerts control over what the latter is able to see.

An important aspect of this relationship is that even a concentrated effort to override the unconscious mind cannot change what happens on the conscious level. Attempting not to be fooled by an optical illusion does not change our visual perception of the image; the only way to see that two distinct-looking colors are indeed the same shade of gray is to remove the surrounding visual context that creates the illusion.

This point brings up an important question regarding the meaning of conscious cognition, said Greenwald.

"Does conscious mean *in* control or *under* control? It actually means both, and both aspects of control are important," he explained. "But the content of conscious cognition is controlled in ways we do not easily understand and intuit."

Just as the influence of the unconscious mind can lead us to make inaccurate sensory judgments, it also can lead us to infer invalid social judgments. Examples of these social illusions abound, including the false — but common — idea that men are more likely to be instrumental virtuosos than are women, or the also incorrect belief that White people, but not Black people, demonstrate good citizenship.

"The limits of our introspective abilities are greater than we understand in our everyday lives," Greenwald said.

Psychological scientists including (from left) **Yarrow Dunham**, APS Fellow **John T. Jost**, APS Past President **Elizabeth Phelps**, and APS William James Fellow **Anthony Greenwald** gathered to celebrate APS Past President **Mahzarin R. Banaji** at a special symposium at the 2017 APS Annual Convention.

Biases on the Brain

During the 1990s, while Greenwald and Banaji were conducting these groundbreaking studies on the implicit expression of social biases, one of Banaji's colleagues in the Yale University psychology department was examining the implicit expressions of emotional learning and memory from a cognitive-neuroscience perspective. Now a professor of psychology at New York University (NYU), APS Past President Elizabeth Phelps was looking for evidence of the role of the amygdala in learning an aversive response in
threat-conditioning paradigms. This connection had already been established in animal models but was not yet confirmed in humans — a topic Phelps saw as highly relevant to Banaji's work on implicit attitudes.

Weaving the two threads together, Banaji and Phelps teamed up to study the relationship between implicit and explicit racial biases and amygdala activation. They found that subjects with a stronger implicit pro-White bias, as measured by the IAT, tended to have more activation in the amygdala when viewing a Black face than when looking at a White face, indicating that these implicit responses were mediated at least in part by the amygdala.

The study, conducted in 1999, "was one of the first clear examples of social neuroscience that got a lot of attention early on, and it was the beginning of our real collaboration," Phelps said.

Banaji and Phelps continued to work together after Phelps moved to NYU, investigating the neurobiological mechanisms underlying racial bias and their consequences. One such study used a threat-conditioning paradigm to examine the phenomenon of "prepared stimuli" that are theorized to elicit a stronger and more persistent fear-learning response due to our evolutionary history (such as our reaction to a potentially deadly spider versus a harmless butterfly).

They found that subjects responded to racial out-group stimuli similarly to how they responded to prepared stimuli: Fear responses to prepared stimuli and racial out-groups were harder to unlearn during

the extinction phase of a threat conditioning experiment than were responses to harmless stimuli and racial in-groups.

“In other words, you have this learning that a stimulus predicts something negative, and it’s much more sticky — it’s harder to get rid of,” Phelps said. Additionally, the two found that this effect was much smaller, or even entirely absent, for subjects who had dated outside of their race, a behavior that may have altered their perceptions of in-group membership.

“This suggested to us that there may be a preparedness to associate negative outcomes with out-group members, and these negative associations may be harder to change with new information,” Phelps explained.

Banaji and Phelps found a similar pattern when they looked at subjects’ judgments of trust. In a classic trust-game paradigm in which study participants had to make decisions about how much money to share with a partner, the researchers found a correlation between implicit race bias and patterns of sharing behavior. The higher a subject’s pro-White bias (as measured by the IAT), the more they shared with White partners relative to Black partners.

Phelps continues to pursue research focusing on ways to control, diminish, or even eliminate these maladaptive emotional and threat reactions.

Whence Implicit Attitudes?

In parallel to the effort to characterize the nature and neuroscience of implicit attitudes, researchers are also interested in understanding how these biases arise in the first place. Former Banaji student Yarrow Dunham is now the Director of the Social Cognitive Development Lab at Yale University, where he examines implicit attitudes from a developmental perspective.

Implicit attitudes were first theorized to be a product of slow learning, in which the level of bias increases across time from a young age, when a child first understands the category of bias (e.g., race or gender), to adulthood, when that bias is firmly entrenched in a person’s implicit social judgments and attitudes. Measuring the in-group preferences of 6-year-old, 10-year-old, and adult subjects (and, in subsequent research, children as young as 3 to 4 years old), Dunham was surprised to find relatively little change in in-group preference across time from 6 years old to adulthood.

“We seem to be seeing, on average, adult-like implicit attitudes right from the beginning,” he explained.

Dunham and colleagues conducted a study in which they measured the preference among young children for their own “minimal group,” an arbitrary distinction based on a random draw of either a red or blue shirt. Even in this context, with no discernible social value assigned to either group, children preferred their own shirt-color group over the other color to a similar degree that 6-year-olds in the earlier studies preferred their own racial in-group.

“This slow-learning account doesn’t seem to get us very far because initial implicit evaluations seem to privilege the in-group right away,” said Dunham. This led him to posit a modified model dubbed “preparedness then tuning,” in which implicit in-group preferences emerge first and are then refined via

slow learning.

Putting this model to the test, Dunham examined the in-group preferences of young Latino Americans compared with their preferences for both another lower-status racial group (Black Americans) and a higher-status group (White Americans). He found that Latino American children show a typical, stable-across-time pattern of preferring their own group over Black Americans. Their in-group preference also remained relatively stable across time compared with their preference for White Americans — but stable at a value of zero. This indicates that subjects had virtually no preference for their in-group over White Americans, "as if these two factors — status internalization and membership — have, in a sense, cancelled each other out," explained Dunham.

While this lack of in-group preference was somewhat expected for adults, who have had a lifetime to internalize a "White = good" bias, Dunham was surprised to see that the same was true for young children.

"What is remarkable here is not that status matters, but that status matters *as much* to quite young children as it does to considerably older children or adults," he said.

This showed that the "preparedness then tuning" model doesn't quite fit either, in that it fails to account for initial attitudes that seem to immediately integrate membership and status. So it's back to the drawing board again for Dunham — not that he isn't excited about the direction future research is headed.

"I think the really tantalizing question is, What are the specific cues that are powerful enough to get a 4-year-old child to counteract or even reverse the tendency towards implicit in-group preference?" he asked.

Since first arriving on the Ohio State campus from her home country of India, Mahzarin Banaji has helped push the boundaries of our understanding about the nature of implicit cognition. As these three collaborators and former students demonstrate, she has also had an enormous influence on other realms of study and on the directions in which psychological inquiry is moving as a whole.

Banaji describes her choice to seek out Greenwald as a graduate advisor and mentor as "the single most important decision that I made, career-wise," a sentiment many of her former students would profess about Banaji herself, who also garners a great deal of personal affection from her mentees. (Former student APS Fellow John Jost, who chaired the symposium, introduced her as "one of my favorite people in the world.") Reflecting on the past 30 years, which have contained all the highs and lows that inevitably accompany the collaboration of two brilliant — and at times clashing — scientists, Banaji is certain of one thing:

"It has not always been easy, but it has always been worthwhile."

References

Chen, J. M., de Paula Couto, M. C. P, Sacco, A. M., & Dunham, Y. (2017). To be or not to be (black or multiracial or white): Cultural variation in racial boundaries. *Social Psychological and Personality*

Science. Advance online publication. doi:10.1177/1948550617725149

Dunham, Y., Baron, A. S., & Banaji, M. R. (2006). From American city to Japanese village: A crosscultural investigation of implicit race attitudes. *Child Development, 7*, 1268–1281.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2007). Children and social groups: A developmental analysis of implicit consistency among Hispanic-Americans. *Self and Identity, 6*, 238–255.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences, 12*, 248–253.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729–738.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the USA, 108*, 7710–7715.

Stanley, D., Phelps, E. A., & Banaji, M. R. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science, 17*, 164–170.

Stanley, D. A, Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., & Phelps, E. A. (2012). Race and reputation: Perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 367*, 744–753.