Title: Modeling How the Brain Updates its Social Expectations

Keywords: stereotypes, cognitive neuroscience, fMRI

Psychology Background: Social expectations play a key role in forming predictions for the characteristics or behaviors of others. For example, if you are a student from a college-environment in which smoking is uncommon, you might expect other students to not smoke and be surprised if they do. Researchers have extensively documented the regularity with which perceivers use the category membership of others to judge them[1,2]. The act of socially categorizing a person, or stereotyping[1,2], manifests in the inferences and assumptions people make about others based on their social category[2]. In the example above, the college student stereotype is that that they do not smoke.

Stereotypes provide a basis from which people can make predictions. In order for their utility to continue, they must be updated over time as they are violated. Continuing from the example above: should you go to a different college-environment in which smoking is common, after numerous exposures to people smoking (violations of your expectations), you eventually begin to expect other students to smoke (you update your expectations). My study seeks to understand how the brain updates its social expectations.

Neuroscience Background: Because stereotyping relies on categorization, it has been studied in relation to semantic knowledge[2]. Stereotype application may draw on cognitive processes that more generally subserve semantic knowledge about categories[1]; however, research suggests that social stereotypes should be considered to be different from other forms of semantic knowledge, as they may relate closely with people's representations of others' mental states[2]. This is supported by the neural activity associated with both stereotypes and Theory of Mind. Research has implicated the regions medial prefrontal cortex (MPFC)[1,3,4], temporoparietel junction (TPJ)[1,3,4], and precuneus/posterior cingulate cortex[1,2,4] both for creating mental representations and for applying social stereotypes. Social expectations have been associated with activity in the anterior cingulate cortex (ACC)[5]. Furthermore, development of context-specific expectations at a single trial level has been associated with left anterior insula, ACC, and dorsolateral prefrontal cortex (DLFPC)[5]. Thus, for a functional Magnetic Resonance Imaging (fMRI) analysis of stereotype application, MPFC, TPJ, and posterior cingulate cortex are prime targets, and for an fMRI analysis of stereotype update, left anterior insula, ACC, and DLFPC.

FMRI imaging data for neural activity of a cognitive process's neural region or network has been shown to be consistent with that process's corresponding computational model[6]. Thus, I can create a computational model for the cognitive process of updating ones social expectations and use it with model-based fMRI techniques to understand, and identify the neural bases of, the neural computations necessary for this update.

Proposed Experimental Model: I propose a general-purpose three-part event-related fMRI experimental model to study the update of social expectations. A key component of the first and third part of this model is the *Implicit Association Test* (IAT), developed by Mahzarin Banaji[7]. The IAT is a computer-administrated test that uses response times to find the strengths of associations between concepts[7]. *Part 1* uses an IAT to find a participant's ("Judge's") baseline probability distribution (BPD) for a set of possible expectations. Once this part is finished, a computer program analyzes the results to generate a probability distribution for the expectation set incongruent with the BPD. I call the computer-generated distribution the update probability distribution (UPD); it is used for the next part. In this part, (*Part 2)*, the Judge is exposed to numerous events where the distribution of correct social expectations follows the UPD. By providing feedback, this part of the model serves to cause the Judge to update his/her expectations through consistent exposure to

events that violate his/her expectations. For example, if four choices A, B, C, D have a BPD such that B and C have the lowest probabilities, then the UPD will be such that either B or C have the highest probability. If B is chosen, then consistent exposure to events where B is most likely to be correct should cause the judge to eventually primarily predict B. *Part 3* involves administering another IAT, which measures the Judge's new probability distribution (NPD) after his/her expectations have been consistently violated. Because this model studies the update of implicit associations, I have coined the term *Implicit Association Update Test* (IAUT).

Motivation: Asian and Hispanic males are often subjected to stereotypes (see personal statement). These stereotypes associate race with (high or low) academic and professional success. To study this, I will apply the IAUT to the question: *How are stereotypes involving the jobs of Asian and Hispanic males updated over time as new violating information is processed?*

Methods: For all events in this IAUT, Judges (participants) will be instructed to predict the jobs of "Actors". A picture of the Actor's face will be presented on the screen, along with labels for the job choices (e.g., *doctor, janitor, gardener, and researcher*). In Part 1, the Judge will be presented with 2 randomly chosen job choices for every Actor. The Judge will be instructed to make a job prediction as quickly as possible. The distribution of job choices for each race will serve to populate the BPD and UPD. In Part 2, the Judge will be presented with all job choices with a distribution for being correct that follows the UPD; however, a sub-section of the beginning will follow the BPD to avoid causing suspicion. He/she will have a small set time (e.g., 5 sec.) to make a prediction and will be shown the correct answer afterwards. Part 3 will be the same as Part 2, and will produce the NPD. For all events, the order of the labels will be randomized. I will pay Judges for each correct prediction to incentive them to be accurate. As control, I will also run this experiment without financial incentives.

Future Directions: I will try to integrate data collected for the UPD, BPD, and NPD to find a possible mathematical relation between the three. I plan to apply this experiment towards studying various other stereotypes, such as other racial stereotypes (i.e. Black and Asian stereotypes) and gender stereotypes.

Broader Impact: The experimental model suggested will prove useful for studying the common phenomenon in which one wishes to study the distribution for a set of social expectations, its update, and its final value. The computer-generated and user-dependent BPD, UPD, and NPD make this model both powerful and adaptable. Furthermore, advancement in our understanding of how social expectations are updated will help us determine how violations of social norms influence our mental representations of the world we live in. As mentioned in my personal statement, people often hold stereotypes in which Black and Hispanic individuals perform poorly academically. My experiment can provide a first step toward the insight necessary to change this.

---

[1]Banaji, Mahzarin R et al. "Neural Correlates of Stereotype Application." (2009) [2] Contreras et al.. "Dissociable Neural Correlates of Stereotypes and Other Forms of Semantic Knowledge." (Social Cognitive and Affective Neuroscience, 2012) [3]Gabrieli, J D E et al. "An fMRI Study of Violations of Social Expectations: When People Are Not Who We Expect Them to Be." (Elsevier, 2011) [4]Adolphs, Ralph. "The Social Brain: Neural Basis of Social Knowledge." [5]Chang, Luke J, and Alan G Sanfey. "Great Expectations: Neural Computations Underlying the Use of Social Norms in Decision-Making." (Social Cognitive and Affective Neuroscience, 2013) [6]O'Doherty, et al. "Model-Based fMRI and Its Application to Reward Learning and Decision Making." (Annals of the New York Academy of Sciences, 2007) [7]Greenwald, Anthony G et al. "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity." (Journal of Personality and Social Psychology, 2009)