Facing up to stereotypes

Martin N Hebart & Chris I Baker

Our understanding of faces reflects both our perception of their facial features and our social knowledge. This interaction of stereotypes and vision can be observed in brain signals in fusiform gyrus and orbitofrontal cortex.

Walking on the sidewalk on a busy day, we can pass hundreds of faces, each of them unique and constantly changing. A face can tell us much about a person, including who they are, how they feel, even what they are attending to. Our ability to extract this information is supported by a network of brain regions that analyze different features of a face and help us form a coherent percept of the person. But this percept is more than just the physical features of the face: it also reflects our prior knowledge and associations of social categories, or stereotypes. These stereotypes facilitate—and often oversimplify—the way we perceive our social world and may cause separate social categories to become related. For example, faces of one race may be more associated with negative facial expressions[1]. Are these stereotypes an interpretation of an unbiased visual representation of a face, or does our social knowledge shape the way we process the visual information itself?

In this issue of *Nature Neuroscience*, Stolier and Freeman[2] demonstrate interactions between stereotypes and the visual processing of faces. They measured stereotypes as associations between sex, race and emotional state. Using fMRI, they observed brain responses congruent with those stereotypes in both the fusiform gyrus, home of the fusiform face area[3], and the orbitofrontal cortex.

The authors measured the influence of stereotypes on subjective perception by asking participants to rapidly categorize images of faces by sex (male, female), race (black, white, Asian), or emotion (angry, happy) (**Fig. 1**). Participants responded using a computer mouse, dragging the cursor to one of the two options displayed on the screen (such as "Angry" or "Happy"). By measuring subtle deviations in the trajectory of the cursor toward the incorrect option, Stolier and Freeman could determine how stereotypes influenced the perception of a given category. For example, when the person saw a picture of a happy male and chose between angry or happy, and their mouse trajectory deviated momentarily toward angry, that would indicate an association or stereotype between male and angry.

The goal of the study was to relate the subjective perception of faces both to social conceptual knowledge and to brain responses. To enable comparison between these

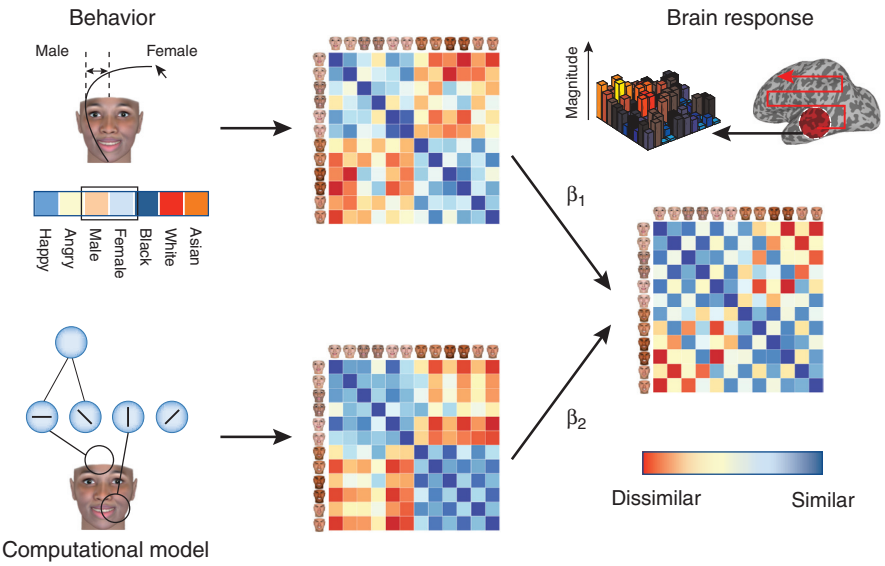

Figure 1 Mapping stereotype responses to brain activity. The authors used dissimilarity matrices to compare behavioral measures of conceptual stereotypes with patterns of brain activity. Participants categorized faces (top left) varying in race, sex and emotional state using a computer mouse. The researchers first generated behavioral dissimilarity matrices by creating a pattern of stereotype responses for each face image (top left) based on deviations in mouse trajectory from the correct response. Then they computed all pairwise comparisons of these patterns and entered them into a matrix (top middle). The faces were also compared with multiple computational visual models (bottom left) to produce dissimilarity matrices based on simple visual features (bottom middle). Finally, fMRI dissimilarity matrices (right) were generated from patterns of brain activity in local 'searchlights' that were moved throughout the entire brain (top right). To identify brain regions reflecting stereotypes while controlling for visual features, the mouse tracking and computational dissimilarity matrices served as predictors for the fMRI dissimilarity matrix in a multiple regression framework. This analysis yielded parameter weights (β_1 and β_2) representing the strength of contribution of each of the predictors to the fMRI responses for each location in the brain. Searchlights showing a match between mouse tracking and fMRI dissimilarity matrices reveal brain regions reflecting the stereotypes held by participants, while controlling for the influence of low-level features.

Martin N. Hebart and Chris I. Baker are in the Laboratory of Brain and Cognition, National Institute of Mental Health, NIH, Bethesda, Maryland, USA. Correspondence should be addressed to M.N.H. martin.hebart@nih.gov or C.I.B. bakerchris@mail.nih.gov

different types of data, the authors formed what is known as dissimilarity matrices. For all possible combinations of sex, race and emotion, these matrices plot the extent to which each pair of faces are related or not—the dissimilarity of those faces (**Fig. 1**). Dissimilarity matrices provide a convenient way to summarize the relative responses to the set of faces and enable a principled approach for comparing representations measured behaviorally with brain activity measured with functional magnetic resonance imaging (fMRI)[4].

Using the mouse tracking data, the authors created a unique stereotype pattern for each face that was composed of the behavioral trajectory data of all seven features that were interrogated during the task (**Fig. 1**). This allowed them to compare the different faces in terms of the correlations between those patterns. For example, the behavioral response pattern to a happy Asian male face may exhibit a greater dissimilarity to the response to an angry Asian male than to the response to a happy white female. All pairwise comparisons were then entered into the dissimilarity matrix.

To confirm that the mouse tracking data do reflect social conceptual knowledge, the authors also asked a separate group of participants to rate a large set of traits (for example, aggressive, intelligent) as stereotypical of a particular sex, race, or emotional state and determined the overlap in these stereotypical associations. The results matched those from the mouse tracking data, confirming that the conceptual associations between social categories are reflected in how they are perceived.

In the next step, the authors built dissimilarity matrices based on fMRI responses. Participants viewed the same set of faces while inside the fMRI scanner. They did not make social judgments about the faces, but were asked to view and remember the faces. To identify brain regions reflecting the mouse tracking dissimilarity matrix, the authors used a 'searchlight' analysis. Specifically, for a given location in the brain, the authors analyzed the local pattern of brain responses in a sphere of voxels centered at that location. By comparing the pattern of response elicited by each face, the authors constructed a fMRI dissimilarity matrix. The crucial step of the analysis was comparing the resulting fMRI dissimilarity matrix to the mouse tracking dissimilarity matrix. This process was repeated for all locations in the brain to identify brain regions showing the strongest correspondence between the fMRI and mouse tracking dissimilarity matrices. These analyses revealed a good match between these dissimilarity matrices in both orbitofrontal cortex and fusiform gyrus in the right hemisphere. In other words, the entangled social category knowledge appeared to be directly reflected in these areas.

An alternative interpretation of this result, however, is that the fMRI dissimilarity matrices reflect not social categories, but visual similarities between the faces that correlate with the social category information but that are not required for eliciting the perception of stereotypes. Indeed, parts of early visual cortex also showed a correspondence with the behavioral dissimilarity matrices.

To rule out a contribution of visual properties, the authors conducted a critical second fMRI experiment in a new group of participants. They minimized the contribution of simple visual information by matching the faces for visual features such as overall contrast. Further, they computed additional dissimilarity matrices based directly on visual information (for example, silhouette of each face, pixel intensities) and on a widely used computational model of the visual processing hierarchy (HMAX[5]). To distinguish the unique contribution of each of these dissimilarity matrices to the fMRI responses, they conducted a multiple regression analysis, with both mouse-tracking and visual-model-based dissimilarity matrices serving as separate predictors. Even when taking the visual models into account, this analysis again pointed to right fusiform gyrus and orbitofrontal cortex, and early visual cortex involvement was no longer observed. These results confirm the interaction of social conceptual knowledge and visual processing in the fusiform gyrus.

Such interactions between conceptual knowledge and visual processing are not limited just to faces but may be a general feature of the visual system. Indeed, the encoding and processing of visual information interacts with attention, value and, more generally, task (for example, refs. 6–8).

An open question is how specific these effects of social categories are to the group of participants tested, who likely all had very similar cultural backgrounds and thus pre-existing knowledge. Directly manipulating stereotypes or testing groups of participants with different stereotypes (for example, from different cultures) would provide stronger support for the interaction of social category knowledge and visual processing. In an additional analysis, the authors showed that the fMRI responses reflected idiosyncratic stereotypes in orbitofrontal cortex, but only weakly so in fusiform gyrus, perhaps as a result of the limited variation across participants.

One limitation of the study is that it is unclear at what stage of processing these interactions emerge. One possible explanation of their results is that there are recurrent interactions between fusiform gyrus and orbitofrontal cortex during the presentation of a face. Alternatively, long-term experience could lead to changes in the bottom-up processing of faces that do not depend on ongoing interactions between regions. In addition, the study does not provide insight into the nature of these stereotype representations. The behavioral and neural representational patterns may be the consequence of a representational space with much lower dimensionality. For example, a principal axis in this representation may be the valence associated with a face (positive versus negative). Investigating the nature of these representations is an important question to be addressed in future research.

Despite these open questions and limitations, the study by Stolier and Freeman[2] provides a striking demonstration of the interaction between the bottom-up processing of visual input and conceptual knowledge.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Hugenberg, K. & Bodenhausen, G.V. *Psychol. Sci.* **15**, 342–345 (2004).
2. Stolier, R.M. & Freeman, J.B. *Nat. Neurosci.* **19**, 795–797 (2016).
3. Kanwisher, N. & Yovel, G. *Phil. Trans. R. Soc. Lond. B* **361**, 2109–2128 (2006).
4. Kriegeskorte, N., Mur, M. & Bandettini, P. *Front. Syst. Neurosci.* **4**, 10.3389/neuro.06.004.2008 (2008).
5. Riesenhuber, M. & Poggio, T. *Nat. Neurosci.* **2**, 1019–1025 (1999).
6. Reddy, L., Kanwisher, N.G. & VanRullen, R. *Proc. Natl. Acad. Sci. USA* **106**, 21447–21452 (2009).
7. Harel, A., Kravitz, D.J. & Baker, C.I. *Proc. Natl. Acad. Sci. USA* **111**, E962–E971 (2014).
8. Hickey, C. & Peelen, M.V. *Neuron* **85**, 512–518 (2015).