nature neuroscience

Neural pattern similarity reveals the inherent intersection of social categories

Ryan M Stolier & Jonathan B Freeman

We provide evidence that neural representations of ostensibly unrelated social categories become bound together by their overlapping stereotype associations. While viewing faces, multivoxel representations of gender, race, and emotion categories in the fusiform and orbitofrontal cortices were stereotypically biased and correlated with subjective perceptions. The findings suggest that social-conceptual knowledge can systematically alter the representational structure of social categories at multiple levels of cortical processing, reflecting bias in visual perceptions.

Individuals extract a wealth of information from another's face, including social categories such as sex, race, and emotion. The traditional view is that each social category dimension is represented independently[1,2], which is sensible as social category dimensions are indeed all distinct and orthogonal in reality. Recent research, however, has questioned the independence of social category dimensions, instead arguing that they may be inherently intertwined.

Current computational models posit strong interdependence between social category representations because shared conceptual knowledge (i.e., stereotypes) related to two ostensibly unrelated categories causes them to become fundamentally entangled[3,4]. These models treat the visual perception of social categories as the end result of lower-level face processing and higher-order social cognition, including stereotypes, mutually constraining one another until a 'compromise' is achieved over time. As facial cues activate categories, categories activate related stereotypes; the stereotypes, in turn, constrain initial category activation itself, while categories and stereotypes recurrently pass activation back and forth[3,4]. Accordingly, the extraction of one category dimension (e.g., sex) will activate stereotypical associations that in turn bias the perception of other category dimensions (e.g., race). To the extent that two social categories (e.g., Male and Black) share stereotypical associations (e.g., Aggressive), those social categories themselves may become linked, even down to the level of visual perception[3–5].

Previous behavioral studies have documented how shared stereotypical associations link various categories together and bias perceptual judgments, including sex and race (Black–Male, Asian–Female)[5], sex and emotion (Male–Angry, Female–Happy)[6], and race and emotion (Black–Angry)[7]. For example, Black and Male stereotypes tend to overlap (Aggressive), just as Asian and Female stereotypes tend to overlap (Docile). Accordingly, previous studies have shown that perceptions of Black male and Asian female faces are facilitated, while perceptions of Asian male and Black female faces are impaired[3,5]. However, despite growing behavioral evidence and support from recent computational models[3], the neural basis of any such intertwined structuring of social categories is unknown.

Neural patterns in the fusiform gyrus (FG) have been shown to represent faces[8] and distinguish their social categories[9]. Whereas areas in early visual cortex (EVC) are more responsive to a face's featural information, FG representations are more responsive to a face's categorical distinctions[10]. Although theoretically predicted[3,5], it remains an open question whether stereotypes' entanglement of social categories could impact perceptual representations of a face in the FG.

Numerous studies have implicated the orbitofrontal cortex (OFC) in modulating perceptions of facial and object stimuli by integrating perceptual representations (for example, in the FG) with top-down expectations activated by contextual or associative details[11,12]. The OFC has also been linked to the integration of facial and contextual cues in social categorization[11], the retrieval of person knowledge[13], and the access of implicit stereotypes[14,15]. These studies suggest that the OFC may help integrate a face's social category cues with top-down expectations about those cues, such as those generated by stereotypical associations. The OFC could then possibly help bias FG representations of a face's categories in line with stereotypical associations and contribute to entangled perceptions.

In two functional magnetic resonance imaging (fMRI) studies, subjects passively viewed (**Fig. 1a** and **Supplementary Fig. 1a**) computer-generated faces that varied in gender (Male, Female), race (Black, White, Asian), and emotion (Angry, Happy), resulting in 12 category-combination conditions (**Supplementary Fig. 2b**). Outside the scanner, subjects completed a mouse-tracking task requiring speeded categorizations of all faces by gender, race, and emotion (**Fig. 1b** and **Supplementary Fig. 1b**). Mouse-tracking is a well-validated measure of how multiple social categories activate and resolve over hundreds of milliseconds during real-time categorization. During two-choice categorization tasks (for example, Male vs. Female), deviation in a subject's hand trajectory toward each category response provides an indirect measure of the degree to which that category was activated during perception (**Fig. 1b** and **Supplementary Fig. 1b**). If stereotypes link one category to another (e.g., Black to Male), subjects' perceptions are biased toward that category and, consequently, their hand trajectories deviate toward that category response in mouse-tracking tasks. Thus, this task measured biased similarities between social categories in subjective perceptions (e.g., Black–Male).

To capture inter-category similarities between all category pairs, dissimilarity matrices (DMs) were generated at multiple levels (stereotype knowledge, subjective perceptions, neural patterns) (**Fig. 2a** and **Supplementary Fig. 2**), and their correspondence was assessed using representational similarity analysis (RSA)[16] (see Online Methods). In study 1 ($n = 17$ subjects), the mouse-tracking data were used to construct a 12 × 12 dissimilarity matrix ('subjective DM'), characterizing the similarity or dissimilarity of each condition pair

Department of Psychology, New York University, New York, New York, USA. Correspondence should be addressed to R.M.S. (rystoli@nyu.edu) or J.B.F. (jon.freeman@nyu.edu).

Received 5 November 2015; accepted 31 March 2016; published online 2 May 2016; doi:10.1038/nn.4296

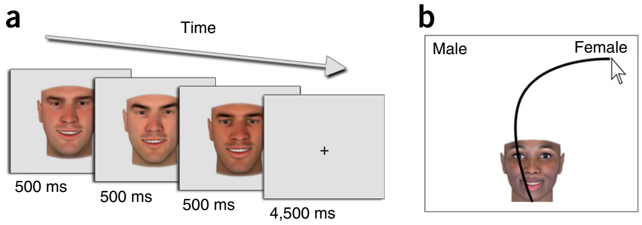

Figure 1 Task designs for study 1 ($n = 17$) and study 2 ($n = 26$). (**a**) Event-related fMRI task. Each encoding event included three exemplars from one condition. Additional probe events ensured participants' visual attention (Online Methods). (**b**) Mouse-tracking behavioral task. On each trial, subjects clicked a start button, a face appeared and they made a categorization response. Hand movement trajectory was recorded en route to the selected response. A trajectory's maximum deviation toward the opposite category response (on the opposite side of the screen) indexed the degree to which that category was activated during perception. In this hypothetical example, a Black female face elicits a trajectory that initially deviates toward the Male response because shared stereotypes between Black and Male categories bias perceptions of Black faces toward male categorization[5]. In our RSA framework, a hypothetical bias to perceive Black faces as more similar to male faces would correspond to a greater similarity between Black and Male categories in the subjective DM.

(for example, similarity of Happy–Asian–Female to Happy–Asian–Male; see Online Methods; **Supplementary Fig. 2b**). A separate stereotype content task, in which each category was assessed for its conceptual relationship with a large set of traits, was conducted on independent raters and used to construct a 7 × 7 'stereotype DM' (**Fig. 2a**). A comparison of similarity values (see Online Methods) within this stereotype DM revealed several biased similarities consistent with prior research (**Supplementary Table 1**), including sex and race (Black–Male, Asian–Female)[5], sex and emotion (Male-Angry, Female-Happy)[6], and race and emotion (Black–Angry)[17]. For behavioral analyses, the 12 × 12 subjective DM (based on condition pairs) was averaged into a 7 × 7 subjective DM based on categories (**Fig. 2a** and **Supplementary Fig. 2c,d**). As hypothesized, RSA confirmed that the subjective DM was significantly predicted by the stereotype DM ($\rho_{(19)} = 0.47$, $P = 0.023$), indicating that conceptual associations between social categories were reflected in how those categories were subjectively perceived (**Fig. 2a** and **Supplementary Fig. 3**).

Our neuroimaging analyses examined the neural basis of such entangled perceptions by performing a whole-brain searchlight RSA throughout each subject's functional data ($P < 0.05$, corrected), identifying cortical regions where the neural-pattern similarity of social categories was predicted by category similarity observed in the mouse-tracking data (subjective DM). As hypothesized, this analysis revealed neural-pattern similarity in the right FG (rFG) and OFC that was significantly predicted by the subjective DM (**Fig. 2b** and **Supplementary Table 2**).

As argued, the biases observed in study 1 may be driven by shared top-down stereotypical associations, an idea that was supported by significant correlation between subjective and stereotype DMs (**Fig. 2a** and **Supplementary Fig. 3**). However, it is also possible that such biases are inherent in the categories' bottom-up visual features themselves. Indeed, significant pattern similarity was additionally observed in the EVC (**Supplementary Table 2**), raising the possibility that these results may be partly explained by low-level visual similarity. In study 2 ($n = 26$), we strengthened our evidence for the role of stereotypical associations by comprehensively controlling for visual similarity and directly measuring each subject's idiosyncratic associations (**Supplementary Figs. 4–6**). Stimuli were matched on low-level visual features (**Supplementary Fig. 4**), and visual similarity was additionally controlled for using three common visual models ('visual DMs'): a hierarchical model of high-level ventral visual representation (HMAX), a model of image silhouettes (retinotopic outlining), and a model of pixel-intensity maps[16] (**Supplementary Fig. 7**) (see Online Methods). To assess subjects' unique stereotypical associations, the stereotype content task was completed by each subject outside the scanner.

Consistent with our hypotheses, participants' subjective DMs were significantly predicted by their own idiosyncratic stereotype DM, even when controlling for all three visual DMs (unstandardized regression coefficient $b = 0.22$, s.e.m. = 0.10, $z = 2.18$, $P = 0.030$), as well as the normative (i.e., group-level) stereotype DM when added to the model ($b = 0.28$, s.e.m. = 0.11, $z = 2.47$, $P = 0.014$; **Supplementary Fig. 8**) (see Online Methods). Thus, biases in how social categories are subjectively perceived reflected subjects' unique conceptual associations about those categories (above and beyond any normative group-level tendency), and they are unlikely to be explained by inherent visual similarity alone.

Whole-brain searchlight multiple regression RSA ($P < 0.05$, corrected) was used to test pattern similarity predicted by each participant's subjective DM while controlling for the three visual DMs. This analysis revealed the rFG (**Fig. 3a**; $x = 27$, $y = -45$, $z = -6$; mean $t = 3.98$; 40 voxels), and when additionally controlling for the normative group-level subjective DM (providing a conservative test that subjects' idiosyncratic biases manifested in neural pattern similarity; **Supplementary Figs. 4** and **6**), the OFC (**Fig. 3b**; $x = -15$, $y = 51$, $z = -27$; mean $t = 3.76$; 33 voxels). These regions therefore contained multi-voxel patterns correlated with subjective perceptions above and beyond inherent visual similarity described by three visual models. As the rFG did not survive the stringent analysis additionally controlling for normative group-level biases at whole-brain correction, given an *a priori* hypothesis we inspected the rFG cluster using a region of interest (ROI) analysis, which confirmed it to be robust to this additional control ($t_{(25)} = 3.12$, $P = 0.005$). These results suggest that a subject's OFC and rFG patterns reflected their own idiosyncratically biased perceptions over and above visual similarity of the stimuli themselves and any normative group-level biases across the sample.

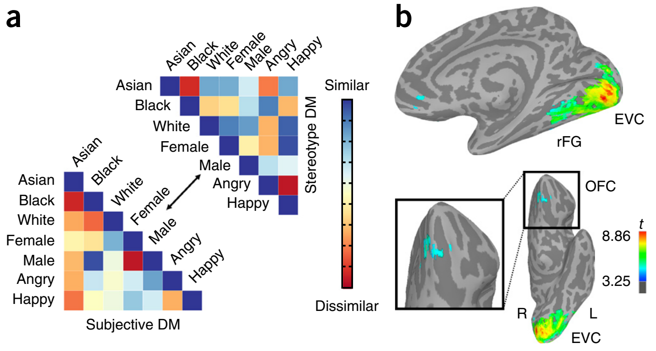

Figure 2 Study 1 results ($n = 17$). (**a**) Diagonal of the subjective DM (in 7 × 7 form) and stereotype DM. Inter-category similarities (Pearson r) in subjective perceptions (subjective DM) were significantly predicted by inter-category similarities in stereotypical associations (stereotype DM), $\rho_{(19)} = 0.47$, $P = 0.023$; correspondence between DMs tested using Spearman correlation RSA). (**b**) Results from searchlight RSA ($P < 0.05$, corrected; voxelwise $P < 0.005$, minimum cluster size, $k > 50$), indicating rFG region ($x = 30$, $y = -55$, $z = -13$; mean $t = 3.80$; 126 voxels) and OFC region ($x = -6$, $y = 37$, $z = -19$; mean $t = 3.95$; 57 voxels) where neural pattern similarity was predicted by the subjective DM (whose pattern similarity was predicted by the stereotype DM in **a** and **Supplementary Fig. 3**). The rFG and OFC regions are depicted on inflated surfaces of the right and left hemispheres, respectively.

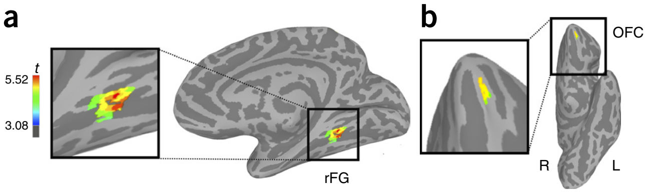

Figure 3 Study 2 fMRI results ($n = 26$). (**a**) Multiple regression searchlight RSA results ($P < 0.05$, corrected; voxelwise $P < 0.005$, $k > 32$) depicted on an inflated surface of right hemisphere, indicating rFG region ($x = 27$, $y = -45$, $z = -6$; mean $t = 3.98$; 40 voxels) where neural pattern similarity was significantly predicted by participants' subjective DMs while controlling for the three visual DMs. An ROI analysis confirmed this region was robust to the additional control of the normative group-level subjective DM; $P = 0.005$. (**b**) Multiple regression searchlight RSA results ($P < 0.05$, corrected; voxelwise $P < 0.005$, $k > 32$) depicted on an inflated surface of left hemisphere, indicating OFC region ($x = -15$, $y = 51$, $z = -27$; mean $t = 3.76$; 33 voxels), where neural pattern similarity was significantly predicted by participants' subjective DMs while controlling for the three visual DMs and the normative group-level subjective DM.

Together the results suggest that rFG and OFC patterns exhibit a representational structure of multiple social categories that is entangled and conformant to stereotypical associations. Behaviorally, we found that the process of perceiving gender, race, or emotion from a face was biased toward how the face, given its other category memberships, was stereotypically expected to appear: namely, Male faces tended to be biased toward Angry categorizations, Female faces toward Happy categorizations, Black toward Male, Asian toward Female, and Black toward Angry, consistent with prior studies[5–7]. Beyond these overall tendencies, subjects' own unique stereotypical associations predicted idiosyncratic biases in their subjective perceptions, and these stereotypically-biased similarities between categories were reflected in the similarity of their multi-voxel representations in the rFG and OFC. Critically, the intertwined nature of these social category representations could not be explained by bottom-up visual similarities; instead, the results suggested top-down social-conceptual knowledge at play that binds seemingly unrelated categories together. These findings confirm recent predictions from computational models[3–5], showing that the brain automatically forms social category representations of a face that are interdependent rather than distinct.

Our results bolster the emerging perspective that higher-order social cognitive processes may dynamically impact lower-level visual processes[3,5,18] by showing that the rFG, a region involved in early face processing[8], contains a representational structure of faces partly shaped by social-conceptual knowledge. For instance, recent studies have found that attitudes[19] and goals[9] alter fusiform representations; here we provide evidence that social-conceptual knowledge, such as stereotypes, also biases such representations. One possibility is that this bias may be imposed by higher-order mechanisms such as the OFC. Indeed, our results are consistent with the perspective that ventral-frontal regions provide top-down perceptual 'priors' in visual cognition[20], such as predictions to facilitate object recognition[12]. Accordingly, the presence of social-conceptual information in the OFC may reflect activation of stereotypes and expectations[14,15,20], which may then sensitize fusiform representations to be in line with those expectations and bias perceptions.

This work is not without its limitations. Although measuring stereotypical associations permitted a test of naturalistic social-category perception, future research could manipulate them to permit stronger causal claims. We should also note that our findings are mute with respect to the origins of the stereotypical associations studied here. In all likelihood, they were acquired across participants' lifespans through cultural transmission and implicit learning[1,2], but it is also possible they were acquired through direct statistical observation of their environments. Regardless of their origin, the results suggest they shape how social categories are subjectively perceived, above and beyond any inherent physical quality of the categories themselves.

In short, our findings suggest that the fundamental structure of social categories when perceiving a face can become warped by social-conceptual knowledge that binds ostensibly unrelated categories together. Thus, although stereotyping has long been considered a consequence of initially perceiving others via categories[1,2], our stereotypes can affect even our initial categorizations. This process is present not only in higher-order regions, such as the OFC, that may play a role in accessing that knowledge, but also in rFG patterns more directly involved in the basic visual processing of another person.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Wagner, J. Haxby, A. Connolly, Z. Ingbretsen, and Y. Xie for insights and assistance in data analysis and D. Amodio and L. Hackel for comments. This work was supported in part by research grant NSF BCS-1423708 (J.B.F.).

AUTHOR CONTRIBUTIONS

R.M.S. and J.B.F. conceived the studies and contributed to study design, analytic strategy, and interpretation of the data. R.M.S. analyzed the data, with the advice of J.B.F. R.M.S. and J.B.F. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fiske, S.T. & Neuberg, S.L. *Adv. Exp. Soc. Psychol.* **23**, 1–74 (1990).
2. Allport, G.W. *The Nature of Prejudice* (Addison-Wesley, Oxford, 1954).
3. Freeman, J.B. & Ambady, N. *Psychol. Rev.* **118**, 247–279 (2011).
4. Freeman, J.B. & Johnson, K.L. *Trends Cogn. Sci.* doi:10.1016/j.tics.2016.03.003 (2016).
5. Johnson, K.L., Freeman, J.B. & Pauker, K. *J. Pers. Soc. Psychol.* **102**, 116–131 (2012).
6. Hess, U., Adams, R.B. Jr. & Kleck, R.E. *Emotion* **4**, 378–388 (2004).
7. Hugenberg, K. & Bodenhausen, G.V. *Psychol. Sci.* **15**, 342–345 (2004).
8. Haxby, J.V., Hoffman, E.A. & Gobbini, M.I. *Trends Cogn. Sci.* **4**, 223–233 (2000).
9. Kaul, C., Ratner, K.G. & Van Bavel, J.J. *Soc. Cogn. Affect. Neurosci.* **8**, 750–755 (2013).
10. Rotshtein, P., Henson, R.N.A., Treves, A., Driver, J. & Dolan, R.J. *Nat. Neurosci.* **8**, 107–113 (2005).
11. Freeman, J.B. *et al. Cereb. Cortex* **25**, 415–422 (2015).
12. Bar, M. *et al. Proc. Natl. Acad. Sci. USA* **103**, 449–454 (2006).
13. Mitchell, J.P., Heatherton, T.F. & Macrae, C.N. *Proc. Natl. Acad. Sci. USA* **99**, 15238–15243 (2002).
14. Milne, E. & Grafman, J. *J. Neurosci.* **21**, RC150 (2001).
15. Knutson, K.M., Mah, L., Manly, C.F. & Grafman, J. *Hum. Brain Mapp.* **28**, 915–930 (2007).
16. Kriegeskorte, N., Mur, M. & Bandettini, P. *Front. Syst. Neurosci.* **2**, 4 (2008).
17. Becker, D.V., Kenrick, D.T., Neuberg, S.L., Blackwell, K.C. & Smith, D.M. *J. Pers. Soc. Psychol.* **92**, 179–190 (2007).
18. Adams, R.B., Ambady, N., Nakayama, K. & Shimojo, S. *The Science of Social Vision* (Oxford University Press, New York, 2011).
19. Brosch, T., Bar-David, E. & Phelps, E.A. *Psychol. Sci.* **24**, 160–166 (2013).
20. Summerfield, C. & Egner, T. *Trends Cogn. Sci.* **13**, 403–409 (2009).

ONLINE METHODS

Analytic approach. Our paradigm measured similarity between social categories at multiple levels. Specifically, for any pair of categories (for example, Black and Male), we aimed to demonstrate that similarity in conceptual knowledge (i.e., stereotypes) related to the two categories predicts similarity in how faces belonging to those categories are subjectively perceived, which in turn predicts the similarity of neural response patterns. For example, we aimed to test whether greater conceptual similarity (i.e., stereotype overlap) between the Black and Male categories predicts greater similarity in how faces belonging to those categories are perceived (i.e., a bias to perceive Black faces similarly to Male faces), which in turn predicts greater similarity in neural response patterns when viewing such faces during fMRI. In study 2, we additionally tested whether these relationships hold when controlling for inherent physical similarity between the categories' are based upon subjects' idiosyncratic biases and stimuli themselves.

In studies 1 and 2, we measured blood oxygenation-level-dependent (BOLD) responses from subjects viewing computer-generated face stimuli in an fMRI scanner. The studies were identical in design and procedure, differing only in two ways. First, study 1 used computer-generated faces and retained their natural appearance; study 2 converted these stimuli into gray-scale and matched the stimulus conditions on luminance and contrast. Second, study 1 collected normative data on the structure of stereotype contents using separate raters; study 2, focusing on subjects' idiosyncratic differences, collected data on the idiosyncratic structure of a subject's own stereotype contents outside the scanner. Face stimuli independently varied along race (Black, White, Asian) × sex (Male, Female) × emotion (Angry, Happy). Because there is no meaningful "neutral" race or sex category, we opted not to include a neutral level of emotion for equivalence. Outside the scanner, subjects completed a series of mouse-tracking categorization tasks, demographics surveys, and in study 2 the additional stereotype content task. The stereotype data were used to predict the mouse-tracking data, which in turn were used to predict the neuroimaging data.

Subjects. In study 1, 17 healthy, right-handed subjects were recruited (47% female; M_{age} = 19.41; 10 white, 6 Asian, 1 Hispanic). Subjects were financially compensated or received partial course credit for participation, and they gave informed consent in a manner approved by the Committee on the Protection of Human Subjects at Dartmouth College. In study 2, a larger sample of 26 healthy, right-handed subjects were recruited (58% female; M_{age} = 20.81; all white). Subjects were financially compensated for their participation, and they gave informed consent in a manner approved by the University Committee on Activities Involving Human Subjects at New York University. The sample size used in study 1 (n = 17) was chosen based on previous fMRI studies using a similar paradigm with RSA methods[21,22] and on fMRI studies examining stereotype effects on face perception[23]. In study 2, given the additional focus on across-subject variability in stereotype structure, the sample size (n = 26) was increased by approximately 50% to permit greater variability and power.

Materials. Face stimuli were generated using FaceGen Modeler. This software uses a 3D morphing algorithm based on anthropometric parameters of the human population[24], in which various social category cues can be precisely manipulated while holding other extraneous cues constant. Faces independently varied along race (Black, White, Asian), sex (Male, Female), and emotion (Angry, Happy). For study 1, 32 unique faces were generated for each condition (for example, 32 Happy Black Female faces), resulting in a total 384 face stimuli. Faces were cropped and vignetted onto a gray background (see **Supplementary Figs. 1** and **2b** for sample stimuli). For study 2, to control for any low-level visual similarity, the stimuli were matched on luminance and contrast across the 12 stimulus conditions using the SHINE toolbox[25] and placed onto a white background (see **Supplementary Fig. 4** for sample stimuli).

Although emotion categories are dynamic, compared with gender or race categories which generally remain static, emotion exhibits categorical perception effects similar to those of static characteristics and is a widely studied form of perceptual categorization[26,27]. Moreover, current models of visually-based social categorization include emotion as a social category[3]. Thus, for purposes of the current work, we refer to emotion as a social category.

Procedure. *fMRI task.* The study procedure in the scanner largely followed that of Connolly *et al.*[21]. The fMRI task was an event-related design over 10 functional runs, where subjects passively viewed faces. There were 6 trials in each run, each trial consisting of: 6 encoding events, 1 null event (fixation), and 1 probe event (recognition). Trials were separated by variable inter-trial intervals (2,000–6,000 ms). The encoding events were presented in a pseudo-randomized order, with the single null event placed between the first and last encoding event, followed by the probe event at the trial end. Probe events were used to ensure subjects maintained attention during face presentations. Each encoding event presented 3 unique face stimuli belonging to a single condition (for example, Happy Black Female) in succession for 500 ms each, followed by a 4,500 ms fixation cross. The null events entailed 6,000 ms of fixation, serving as a measure of baseline. Probe events were identical in timing to encoding events; however, participants viewed either faces already presented within the trial or novel faces to that trial (pseudo-randomized so as to be counterbalanced). Participants were asked to provide a button-press if the faces were presented earlier that trial. Face stimuli within a given encoding event were randomly selected from the condition of that event without replacement until all stimuli were selected. Accordingly, each face was presented 2 or 3 times by the end of the task, depending on the order to which they were selected without replacement until all had been presented. All 12 conditions were presented every 2 trials. Altogether, each condition was presented during 30 separate encoding events.

Mouse-tracking tasks. Subjects completed mouse-tracking tasks to assess how gender, race, and emotion categories biased one another in subjective perceptions and to compute dissimilarity matrices (DM) used for neuroimaging analyses. The mouse-tracking data were collected using the well-validated MouseTracker software[28]. Standard two-choice tasks were implemented. In a randomized order, subjects completed 5 mouse-tracking tasks, in which they categorized target stimuli along Sex (Male vs. Female), Emotion (Angry vs. Happy), or Race (White vs. Black, Black vs. Asian, White vs. Asian). All face stimuli were presented once in the Sex and Emotion tasks. Each face stimulus was presented in only the 2 applicable (out of 3) race tasks. For example, a Happy Black Female stimulus would be presented in the Black vs. White and Black vs. Asian tasks, but not in the White vs. Asian task. In all 5 tasks, to begin each trial subjects clicked on a 'Start' button located at the bottom-center of the screen, which was immediately replaced by the face stimulus. Face stimuli were presented in a randomized order. Subjects were asked to categorize the stimulus as quickly and accurately as possible by selecting one of two responses (for example, Male vs. Female, Asian vs. Black), located in the top-left and top-right corners of the screen, using a mouse-click. The left/right location of responses was counterbalanced across subjects. During the categorization process, the streaming x, y coordinates of the mouse were recorded (sampling rate ~70 Hz). To ensure trajectories were on-line with the actual decision process, we encouraged participants to begin initiating movement as early as possible. As in previous research, if movement initiation time exceeded 400 ms, a message appeared after subjects made their response, encouraging them to start moving earlier on future trials even if they were not fully certain of their response. Further details on the methodology and analytic techniques for mouse-trajectory data can be found elsewhere[28]. See **Figure 1** and **Supplementary Figure 1** for a schematic illustration of the process.

Stereotype content task. To measure stereotype contents (i.e., conceptual knowledge) associated with each social category, we collected ratings on a large set of traits for gender, race, and emotion categories. In study 1, these data were collected from a separate group of raters to provide normative information about stereotype content. This allowed us to assess the degree to which common stereotypical associations specifically underlie the biased similarity structure observed in subjective perceptions (the subjective DM, from mouse-tracking data). The subjective DMs, in turn, were used to predict neural patterns. Raters in study 1 were recruited through Amazon Mechanical Turk online and received monetary compensation: gender task (n = 20; 55% female; M_{age} = 32.65; 16 White, 2 Asian, 1 Black, 1 American Indian), race task (n = 18; 39% female; M_{age} = 37.06; 14 White, 1 Black, 2 Hispanic, 1 other), and emotion task (n = 21; 48% female; M_{age} = 32.95; 17 White, 2 Black, 1 Asian, 1 Other). Five raters were excluded because they did not correctly follow the instructions. Following the survey, raters also completed a standard demographics questionnaire. In the task, raters were presented with 96 traits (for example, 'aggressive', 'intelligent') one at a time in a randomized order and indicated whether the trait was stereotypic of the Male or Female category (gender task); Black, White, or Asian category (race task); or Angry or Happy category (emotion task); using dichotomous judgments (yes or no). Ratings of each category were blocked; the order of blocks (for example, Male first, Female second) was counterbalanced across raters. The traits included were a comprehensive set used in previous work documenting overlapping Black–Male and Asian–Female stereotypical associations and several additional traits[29–31]. To prevent social desirability concerns from biasing responses, raters were instructed to answer based on what they believed the held stereotypes are of a typical American.

In study 2, given the focus on idiosyncratic differences among subjects, the same subjects who underwent fMRI scanning completed the stereotype content task following the mouse-tracking task, in which they rated all 7 categories on 96 traits. This allowed us to explore the extent to which the unique structure of a subject's stereotype contents was present in their subjective perceptions (subjective DM), which in turn was used to predict neural patterns. To permit more precise assessment using a wider range of options, the 96 traits in study 2 were rated on a 7-point scale from "non-stereotypic" to "very stereotypic".

fMRI acquisition and preprocessing. In study 1, subjects were scanned using a 3T Philips Intera Achieva scanner in the Dartmouth Brain Imaging Center. Anatomical images were acquired using a T1-weighted protocol (MPRAGE, 1.33 × 1.33 × 1.33 mm). Functional images were acquired using a single-shot gradient echo EPI sequence (TR = 2,000 ms, TE = 35 ms) using 35 interleaved oblique-axial slices (3 × 3 × 4 mm voxels; no slice gap). In study 2, subjects were scanned using a 3T Siemens Allegra scanner in the NYU Center for Brain Imaging. T1-weighted anatomical images were collected (MPRAGE, 1 × 1 × 1 mm), and functional images using 34 slices (3 × 3 × 3 mm voxels; no slice gap) were acquired using a customized multi-echo EPI sequence (TR = 2,000 ms, TE = 30 ms) developed by the NYU Center for Brain Imaging, which was optimized for mitigating susceptibility artifacts in the OFC and medial-temporal regions. In both studies, preprocessing of the imaging data was conducted using the most recent version (16.0.09) of AFNI software[32]. Functional imaging data preprocessing included high-pass filtering of frequencies, slice timing correction, 3D motion correction, voxelwise detrending, and spatial smoothing using a 3D Gaussian filter (4-mm FWHM). Structural and functional data of each subject were transformed to standard MNI space.

Mouse-trajectory preprocessing. Standard mouse-tracking preprocessing was used[28]. All response trajectories were rescaled into a standard coordinate space (top left: [−1.0, 1.5]; bottom right: [1.0, 0]) and normalized into 100 time bins using linear interpolation to permit averaging of their full length across multiple trials. For comparison, all trajectories were remapped rightward. To obtain a by-trial index of the degree to which subjects were biased to select the incorrect gender, race, or emotion category (on the opposite side of the screen), we calculated the maximum deviation (MD) of each mouse trajectory toward the opposite response option. MD is measured by the maximum perpendicular deviation from an idealized straight-line trajectory between its start and endpoints. During two-choice mouse-tracking categorization tasks (for example, Male vs. Female), deviation in a subject's mouse trajectory toward an opposite category response (indexed by MD) is a well-validated measure of the degree to which that category was activated during the perceptual process (**Fig. 1b** and **Supplementary Fig. 1b**)[28,33,34]. If stereotypical associations link one category to another (for example, Black to Male), subjects' perceptions are systematically biased toward that category and, consequently, hand trajectories deviate toward that category response in mouse-tracking paradigms[5]. The pattern of average MD values in the various conditions was used to develop dissimilarity matrices (DMs), indexing the similarity or dissimilarity of each face category (for example, Black) to another category (for example, Male) in subjective perceptions. Such DMs were then used to predict neuroimaging data, as detailed in the sections that follow.

Pattern similarity analyses preparation. For each subject, searchlight representational similarity analysis (RSA) requires a single whole-brain pattern of activation per condition. These patterns were used to assess representational similarity between condition pairs. Accordingly, we estimated the average hemodynamic response for each condition at every voxel and for every run using the 3dDeconvolve procedure in AFNI. Specifically, subjects' BOLD responses were modeled using a GLM whose design matrix included a total of 13 predictors: 12 predictors for each of the 12 stimulus conditions, and one additional predictor to model probe events. All predictors were modeled as boxcar functions across the first 2 s of each event (during which the face stimuli were presented) and convolved with a gamma variate function (GAM in AFNI). We used the resulting voxelwise t statistics comparing condition responses to baseline to comprise the whole-brain patterns of activation for each of the 12 conditions. These statistics were then averaged across runs at every voxel and used for RSA. Neural responses associated with probe events were not used for subsequent analyses.

Overview of pattern analyses and dissimilarity matrices (DMs). Dissimilarity matrices (DMs) are the symmetrical matrices of dissimilarity between all condition-pairs (conditions × conditions; in our case, 12 × 12 category combinations). In such a matrix, larger values represent larger dissimilarity of pairs, such that the smallest value possible is the similarity of a condition unto itself (dissimilarity of 0). Dissimilarities between conditions used to construct the stereotype, subjective, visual, and neural DMs were computed as Pearson correlation (study 1) or squared Euclidean (study 2) distances[16] (details on these various DMs are provided below). To assess the correspondence between neural data (neural DMs) with our models (subjective and visual DMs), DMs were transformed into vectors of their unique values (values under the diagonal), and the Spearman rank-order correlation between these dissimilarity vectors was assessed (rank-ordering is preferable when comparing DMs from different measures as it does not assume a linear relation)[16,35].

Searchlight RSA was completed using PyMVPA[36]. The searchlight RSA technique performs a searchlight[37] across the brain seeking regions where the similarity of the neural voxel-patterns per condition significantly correlates with that predicted by a model. Specifically, within a 3-voxel radius (123 voxels) sphere defined around each searchlight center voxel, the Spearman correlation between neural representational-similarity (specifically, the neural DM) and that predicted by a model based on behavioral or stimulus-derived data (for example, subjective or visual DM) was computed. Spearman ρ values were then mapped back upon the center voxel, yielding a whole-brain map showing the representational similarity of the conditions between the predicted model and neural data per subject. Specifically, the procedure mapped Spearman ρ coefficients to each voxel per subject, indexing the correlation between the vector-form distances (flattened unique values under the diagonal of the DM) of the subjective DM and neural DM within that voxel's respective searchlight sphere. To test this relationship for significance, the Spearman ρ values were then Fisher-z transformed and t-tested at a group level against zero (using AFNI's 3dttest++ program).

In study 1, this RSA was used to predict neural similarity from the average subjective DM, mapping Spearman ρ correlation coefficients back to each searchlight sphere center. In study 2, we used multiple regression analyses (ordinary least-squares) rather than correlation to additionally control for alternate predictor DMs. These analyses therefore performed the same procedure, while also controlling for the relationships between a number of alternate DMs (for example, visual DMs) and the local sphere's neural DM. Thus, regression beta coefficients (rather than correlation coefficients) were mapped to each voxel per subject. The beta coefficients were then t-tested at a group level against zero (using AFNI's 3dttest++ program). In both studies, we corrected for multiple comparisons (false positive rate < 0.05) using Monte Carlo simulations (3dClustSim in AFNI; smoothness estimated with a spatial autocorrelation function). Simulations demonstrated that an experiment-wide $\alpha < 0.05$ was maintained using a voxelwise threshold of $P < 0.005$ and a minimum cluster extent of 50 (study 1) and 32 (study 2) voxels.

In study 2, we used multiple regression RSA to predict neural similarity with a subject's unique subjective DM over and above other models of category similarity. Because multiple regression RSA assumes a linear combination of multiple predictor DMs, these analyses require a similarity measure that sums linearly; thus, we used squared Euclidean distances[38]. Prior to computing the measure, pattern vectors were z-normalized in order to isolate the relative pattern of each condition (removing absolute differences in vector magnitude and scale), which is sensible given our Pearson correlation distance approach to pattern similarity in which normalization is inherently carried out by a correlation distance[39,40]. Indeed, squared Euclidean distances of normalized pattern vectors are equivalent (i.e., linearly proportional) to correlation distances[41], and we therefore report statistics using correlation distances for ease of understanding and consistency with study 1.

In study 2, we first performed multiple regression RSA, regressing neural-pattern similarity on the subjective DM while simultaneously controlling for three visual DMs (HMAX model, image silhouettes, pixel-intensity maps; see

"Visual DMs," below). Together with the matching of stimuli on low-level visual properties in study 2, controlling for visual DMs provides a stronger test of the unique contribution of social-conceptual knowledge on rFG and OFC representation structure above and beyond any possible inherent visual similarity. Note that these visual DMs additionally control for neural patterning that might merely reflect the 12 unique category combinations or 3 unique category dimensions (gender, race, emotion) without any systematic biases. We added the normative group-level subjective DM to the model in order to provide a more conservative test showing that stereotype-driven category similarity unique to each subject was reflected in that subject's neural-pattern similarity (controlling for the normative structure of stereotypic associations in the sample). Because the significant rFG cluster elicited by the analysis controlling for visual models did not survive the conservative analysis of additionally controlling for the normative group-level subjective DM at whole-brain correction, given an *a priori* hypothesis we inspected the rFG cluster using an ROI analysis to confirm its robustness to the additional control. The RSA effect (beta values) in the rFG cluster (40 voxels) was tested against 0 using a one-sample *t*-test. Note that the critical test of the rFG effect's robustness to visual controls is guaranteed by the independent whole-brain analysis.

Stereotype DM. Using data from the stereotype contents task, a stereotype DM was created to assess the similarity or dissimilarity in conceptual associations of the various social categories based on either data from separate raters (study 1) or idiosyncratic data from each individual scanner subject (study 2). For study 1, we averaged responses across independent raters within each trait and target category, creating a proportion score per trait and per category. We then computed the pairwise Pearson correlational distance between all categories in the stereotype contents data (1 – *r* between each of the 7 category vectors of ratings on the 96 traits). This produced a 7 × 7 stereotype DM between the 7 categories to assess those categories' similarity in stereotype contents (Pearson correlational distance; see **Fig. 2a**). In study 2, idiosyncratic stereotype DMs were created for each scanner subject. As in study 1, we calculated the Pearson correlation distance between all categories in their stereotype contents data (1 – *r* between each of the 7 category vectors of ratings on the 96 traits) within each subject. This produced a 7 × 7 stereotype DM between the 7 categories unique to each subject. A normative group-level stereotype DM was also computed as the mean 7 × 7 DM of all idiosyncratic stereotype DMs to use as a statistical control (**Supplementary Figs. 4** and **6**).

Because the stereotype DM exists only in 7 × 7 form (based on categories) rather than 12 × 12 form (based on category combinations), it is not possible to correlate it with a searchlight sphere's neural DM (in 12 × 12 form) directly. Moreover, our hypotheses concern neural regions that are involved in representing subjective perceptions (assessed via mouse-tracking), which may in turn may be biased by stereotypic associations. Accordingly, the relationship between neural DMs and stereotype DMs is necessarily indirect. Thus, our analytic approach was to first collapse 12 × 12 subjective DMs into their averaged 7 × 7 form, and to predict such 7 × 7 subjective DMs from stereotype DMs (in study 2, while additionally controlling for a number of alternate model DMs). Establishing that stereotype structure is present in subjective perceptions (subjective DMs), we then aimed to predict neural DMs (12 × 12) from subjective DMs (which were shown to reflect stereotype structure) in their original 12 × 12 form (in study 2, while additionally controlling for a number of alternate model DMs).

Subjective DM. Using data from the post-scan mouse-tracking tasks, a subjective DM for each subject was created to assess the similarity or dissimilarity of the various social categories at the level of subjective perceptions. In study 1, data from one subject were excluded due to a data-recording malfunction. Incorrect categorization responses and trials with reaction times exceeding 2,000 ms were excluded. As described earlier, during two-choice mouse-tracking categorization tasks (for example, Male vs. Female), deviation in a subject's mouse trajectory toward each category response (indexed by MD) provides an indirect measure of the degree to which that category was activated during perception (**Fig. 1b**)[28,33,34]. The pattern of average MD values in the various conditions was used to develop the subjective DMs (see earlier for more information on the MD measure). For example, the extent to which subjects were biased to select 'Male' while categorizing Black faces can be conceptualized as higher similarity between Male and Black categories. Such stereotypically biased similarity between Male and Black would therefore be reflected in the subjective DM, which in turn was used to predict neural patterns showing a correspondingly biased similarity.

For each of the 12 stimulus conditions, we treated the average MD relative to the maximum possible MD [MD/max(MD)] as similarity toward the unselected response on the opposite side of the screen (incorrect response; for example, Happy Black Female similarity to 'White'), and inverse effect [1 – (MD/max(MD))] as similarity toward the selected response (correct response; for example, Happy Black Female similarity to 'Black'). In such a way, the physical proximity or distance of the trajectory toward the unselected versus selected responses served as a proxy for the similarity or dissimilarity of the particular stimulus condition to the 7 response categories across the mouse-tracking categorization tasks. Accordingly, each task provided similarity measurements of each of the 12 conditions categorized to both response options in that task (Emotion task: 'Angry' versus 'Happy'; Sex task: 'Female' versus 'Male'; Race task a: 'Asian' versus 'Black'; Race task b: 'Asian' versus 'White'; Race task c: 'Black' versus 'White'). For each of the 12 stimulus conditions, the result was a vector of 1 similarity value toward each of the 7 basic category responses ('Asian', 'Black', 'White', 'Female', 'Male', 'Angry', 'Happy') (**Supplementary Fig. 2a**). We then computed the final condition-pair dissimilarities as the Pearson correlational distance between their respective 7-length category-similarity vectors (**Supplementary Fig. 2a**), resulting in a 12 × 12 subjective DM for each subject (**Supplementary Figs. 2b**, **4** and **6**). As such, similarity in this context means activating the 7 response categories in a similar fashion during categorization. The subjective DM thus captures the extent to which various social categories are biased toward one another in subjective perceptions.

Visual DMs. As we are interested in biased category similarity due to conceptual knowledge rather than any inherent visual similarity from mere physical resemblance, in study 2 we not only matched stimuli on several visual characteristics, but we also developed visual DMs to control for possible effects from physical resemblance (**Supplementary Fig. 7**). We computed multiple DMs based on several models representing the similarity or dissimilarity in visual features composing the 12 stimulus conditions. Accordingly, these visual DMs captured inherent physical resemblance between each pair of category conditions. Specifically, we selected three visual models used in previous research to account for different levels of visual representation relevant to our stimuli[16].

HMAX C2. A visual DM was created from the C2 layer of the HMAX model of ventral-visual stream representation[42,43]. This model simulates higher-level representation in the object recognition visual stream, meant to closely resemble representation in extrastriate visual area V4 or posterior inferior temporal cortex (IT). This model has been found to best account for similarity structures in the FG compared with other visual computational models[16]. HMAX features per stimulus were estimated, and averaged within face category conditions. These feature vectors were then correlated for each condition-pair to generate a final HMAX visual DM.

Image silhouette. To model low-level visual properties of the images represented in EVC, we computed a silhouette DM, which was the condition-pair correlations between average flattened pixel-intensity maps of face stimuli that were transformed into silhouettes (the entire face image converted to 0s, background to 1s). Silhouette models have been found previously to best account for representational similarity in EVC, even more so than models of specific regional cell types[16]. Additionally, one concern with matching these stimuli on low-level features was that they could not be matched on retinotopic outlining due to facial shape cues playing an important role in categorization[44]. The silhouette model captures retinotopic outlining of the images and therefore serves as a control of this feature.

Pixel-intensity maps. While the silhouette model has been found to account well for EVC representation in previous work, it of course neglects many other low-level image properties besides image outlines. We therefore included a model of the general image similarities to account for other low-level visual features of the images. Specifically, a luminance pattern DM was created by correlating all pairs of the average flattened pixel-intensity maps per condition.

Analysis of behavioral data. In both studies, to examine the nature of social category similarity in subjective perceptions (the subjective DM), we broke down each subject's 12 × 12 subjective DM (**Supplementary Fig. 2b**) into

7 × 7 DMs corresponding to the 7 basic categories (Male, Female, Black, White, Asian, Angry, Happy) (**Fig. 2a** and **Supplementary Fig. 2c,d**). We could then use this 7 × 7 DM to test for the presence of stereotype-driven category similarities. The collapsing of the subjective DM into a 7 × 7 format of basic categories was necessary to compare it to the 7 × 7 stereotype DM.

In the original 12 × 12 DMs, each of the 12 conditions (for example, Happy Black Male) had an associated vector of response activation (indexed by maximum deviation (MD)) for the 7 basic categories (for example, Angry), and dissimilarity was computed on the basis of those vectors. To compute the 7 × 7 DM, we performed a procedure similar to that used to create the 12 × 12 subjective DM. First, for each of the 7 basic categories, we created a vector of similarity to each category response option from the mouse-tracking data (for example, Male similarity to Asian, Black, White, Female, Male, Angry, Happy), by averaging all of the aforementioned response activation vectors (of each condition's similarity to the 7 basic category response options) of conditions containing that category. For example, the Male vector of response activation was the average similarity of the Happy Black Male, Happy White Male, Happy Asian Male, Angry Black Male, Angry White Male, and Angry Asian Male conditions to the 7 category response options (such as how an average of the vectors in **Supplementary Fig. 2a** would index average Asian similarity to the 7 basic category responses). The pairwise Pearson correlation distance between these vectors created a 7 × 7 subjective DM (**Fig. 2a**), indexing similarity between the 7 basic categories (for example, distance of Male to Female). Incorrect categorization responses and trials with reaction times exceeding 2,000 ms were excluded.

It should be noted that, while the averaging procedure to derive the 7 × 7 DM from the 12 × 12 DM does involve overlapping observations used to compute different categories, and these overlaps produce covariance between categories (for example, Black is calculated from Black Males, and Male is calculated from Black Males), there is an equal amount of dependence between non-mutually exclusive categories (for example, Black is calculated from an equal number of Black Males and Black Females). That is, the inherent similarity between non-mutually exclusive categories (for example, Black to Male) due to overlapping observations is uniform between categories (for example, Black is inherently as related to Male and Female). Furthermore, this procedure does not produce inherent dissimilarity within categories (for example, sex), as they do not share any observation and are thus independent (for example, Male is not computed with any Female observations). Since the purpose of this analysis is to investigate interdependence between social category dimensions (for example, sex and race, such as Male to Black versus Asian), and not within them (for example, Male versus Female), this has no confounding effect on any of the analyses.

In study 1, to first test similarity between categories' stereotypic associations (see **Supplementary Table 1**), we compared Pearson *r* coefficients (extracted from the stereotype DM) by Fisher *z*-transforming them, then computing a *z* score between them (allowing us to significance test against a standard normal distribution[45]). Then, to assess the degree to which such overlapping stereotypes predicted biased similarities in subjective perceptions (mouse-tracking data), we conducted RSA[16] between the stereotype overlap data (7 × 7 stereotype DM) and mouse-tracking data (7 × 7 subjective DM). Specifically, we tested the Spearman correlation between the unique inter-category similarities under the diagonal of the stereotype DM and subjective DM (each therefore with 21 observations).

While study 1 used normative stereotype data from separate raters, study 2 examined subjects' idiosyncratic differences in stereotypic associations (which, in turn, were used to predict representational similarity in subjective perceptions and neural patterns). Accordingly, **Supplementary Figure 5** depicts subjects' variability in inter-category similarities in stereotypic associations (stereotype DMs) and subjective perceptions (subjective DMs). Using multiple regression RSA, we tested whether subjects' idiosyncratic stereotype DMs predicted their subjective DMs while controlling for possible effects of inherent visual similarity (three visual DMs of HMAX output, image silhouettes, and pixel-intensity maps) as well as any overall normative group-level tendencies (the normative group-level subjective DM is depicted in **Supplementary Figs. 4** and **6**). Due to the need to examine idiosyncratic subject-specific effects using multiple predictor DMs (unlike behavioral analyses of study 1), we used a multi-level regression framework with generalized estimating equations (GEE) in order to appropriately account for the correlated nature of repeated measurements within subjects[46]. No assumptions were made about the specific correlation structure a priori (unstructured correlation matrix). Unstandardized regression coefficients, standard errors, and Wald *z* statistics are reported.

Relationship between subjective DM and visual DMs. To confirm that the stimulus matching on low-level visual properties of study 2 was effective in reducing stimulus confounds in study 1, a multiple regression RSA (using ordinary least-squares) was used to estimate the relationship between the three visual models (visual DMs) and the subjective DM in studies 1 and 2. We anticipated that visual models would be significantly correlated in both studies, because even if perceptions of facial stimuli (subjective DM) are partly biased by social-conceptual knowledge, as we predict, they clearly should still reflect the stimuli's visual characteristics (visual DMs) to some sizable degree as well. Thus, the subjective DM should likely be explained by the visual DMs regardless of how stimuli are matched, as the physical cues providing the basis for categorization will always affect mouse-tracking response trajectories during categorization (which in turn are used to compute the subjective DM). However, we anticipated considerably stronger effects of the visual DMs on the subjective DM in study 1 than study 2. First, we regressed the study 1 12 × 12 subjective DM onto the three 12 × 12 predictor DMs of the study 1 stimuli (HMAX output, image silhouettes, and pixel intensity maps) using multiple regression RSA (unique values under the diagonal resulted in 66 observations for each DM). Variables were *z*-normalized before analysis. The combined omnibus effect of the visual DMs on the subjective DM was quite strong in study 1 (mean $b = 0.34$, $F_{(3,62)} = 59.86$, $P < 0.0001$). As expected, conducting the analogous analysis in study 2 (regressing the study 2 subjective DM onto the three visual DMs of study 2 facial stimuli) indicated a notably weaker effect of the visual DMs on the subjective DM (mean $b = 0.18$, $F_{(3,62)} = 18.14$, $P < 0.0001$). These results confirmed that low-level visual properties matching in study 2 was successful in reducing stimulus confounds in study 1.

Statistics. All tests were two-tailed. For RSA, standard techniques were used. Spearman rank order correlation analyses were used to assess relationships between two DMs. When controlling for additional predictor DMs, we used ordinary least-squares multiple regression. Given the need to assess subject-specific correspondence between stereotype and subjective DMs in study 2, RSA of study 2 behavioral data was performed in a multi-level regression framework using GEE in order to appropriately account for the correlated nature of repeated measurements within subjects (no assumptions were made about the correlation structure). In order to compare similarity values within the stereotype DM of study 1, we compared the cell Pearson *r* coefficients by Fisher *z*-transforming them, then computing a *z*-score between them (allowing us to significance test against a standard normal distribution[45]). In all analyses, although not formally tested, data distributions were assumed to be normal. Both studies were within-subjects designs involving no group allocation; therefore, blinding to any between-subject conditions and randomization to such conditions was not applicable.

A **Supplementary Methods Checklist** is available with additional details about all reported analyses**.**

Data and code availability. The data that support the findings of this study are available from the corresponding authors upon request. The code used for the analyses also is available upon request.

21. Connolly, A.C. *et al. J. Neurosci.* **32**, 2608–2618 (2012).
22. Kietzmann, T.C., Swisher, J.D., König, P. & Tong, F. *J. Neurosci.* **32**, 11763–11772 (2012).
23. Hehman, E., Ingbretsen, Z.A. & Freeman, J.B. *Neuroimage* **101**, 704–711 (2014).
24. Blanz, V. & Vetter, T. *SIGGRAPH'99* 187–194 (ACM Press, Los Angeles, 1999).
25. Willenbockel, V. *et al. Behav. Res. Methods* **42**, 671–684 (2010).
26. Calder, A.J., Young, A.W., Perrett, D.I., Etcoff, N.L. & Rowland, D. *Vis. Cogn.* **3**, 81–118 (1996).
27. Etcoff, N.L. & Magee, J.J. *Cognition* **44**, 227–240 (1992).
28. Freeman, J.B. & Ambady, N. *Behav. Res. Methods* **42**, 226–241 (2010).
29. Galinsky, A.D., Hall, E.V. & Cuddy, A.J. *Psychol. Sci.* **24**, 498–506 (2013).

30. Devine, P.G. & Elliot, A.J. *J. Pers. Soc. Psychol.* **21**, 1139–1150 (1995).
31. Katz, D. & Braly, K. *J. Abnorm. & Soc. Psych.* **28**, 280–290 (1933).
32. Cox, R.W. *Comput. Biomed. Res.* **29**, 162–173 (1996).
33. Freeman, J.B., Dale, R. & Farmer, T.A. *Front. Psychol.* **2**, 59 (2011).
34. Spivey, M.J. & Dale, R. *Curr. Dir. Psychol. Sci.* **15**, 207–211 (2006).
35. Carlin, J.D., Calder, A.J., Kriegeskorte, N., Nili, H. & Rowe, J.B. *Curr. Biol.* **21**, 1817–1821 (2011).
36. Hanke, M. *et al. Front. Neuroinform.* **3**, 3 (2009).
37. Kriegeskorte, N., Goebel, R. & Bandettini, P. *Proc. Natl. Acad. Sci. USA* **103**, 3863–3868 (2006).
38. Carlin, J.D. & Kriegeskorte, N. Preprint at doi:10.1101/029603 (2015).
39. Khaligh-Razavi, S.-M. & Kriegeskorte, N. *PLoS Comput. Biol.* **10**, e1003915 (2014).
40. Alink, A., Walther, A., Krugliak, A., van den Bosch, J.J. & Kriegeskorte, N. Preprint at doi:10.1101/032391 (2015).
41. Nili, H. *et al. PLoS Comput. Biol.* **10**, e1003553 (2014).
42. Serre, T., Wolf, L. & Poggio, T. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005* **2**, 994–1000 (IEEE, 2005).
43. Riesenhuber, M. & Poggio, T. *Nat. Neurosci.* **2**, 1019–1025 (1999).
44. Freeman, J.B. & Ambady, N. *Psychon. Bull. Rev.* **18**, 705–712 (2011).
45. Rosenthal, R. & Rosnow, R.L. *Essentials of Behavioral Research: Methods and Data Analysis* (McGraw-Hill, New York, 2007).
46. Burton, P., Gurrin, L. & Sly, P. *Stat. Med.* **17**, 1261–1291 (1998).