PNAS RESEARCH ARTICLE | NEUROSCIENCE
PSYCHOLOGICAL AND COGNITIVE SCIENCES

Neural representations of others' traits predict social decisions

Kenji Kobayashi[a,1], Joseph W. Kable[a], Ming Hsu[b], and Adrianna C. Jenkins[a,1]

Edited by Elke Weber, Princeton University, Princeton, NJ; received September 14, 2021; accepted March 31, 2022

To guide social interaction, people often rely on expectations about the traits of other people, based on markers of social group membership (i.e., stereotypes). Although the influence of stereotypes on social behavior is widespread, key questions remain about how traits inferred from social-group membership are instantiated in the brain and incorporated into neural computations that guide social behavior. Here, we show that the human lateral orbitofrontal cortex (OFC) represents the content of stereotypes about members of different social groups in the service of social decision-making. During functional MRI scanning, participants decided how to distribute resources across themselves and members of a variety of social groups in a modified Dictator Game. Behaviorally, we replicated our recent finding that inferences about others' traits, captured by a two-dimensional framework of stereotype content (warmth and competence), had dissociable effects on participants' monetary-allocation choices: recipients' warmth increased participants' aversion to advantageous inequity (i.e., earning more than recipients), and recipients' competence increased participants' aversion to disadvantageous inequity (i.e., earning less than recipients). Neurally, representational similarity analysis revealed that others' traits in the two-dimensional space were represented in the temporoparietal junction and superior temporal sulcus, two regions associated with mentalizing, and in the lateral OFC, known to represent inferred features of a decision context outside the social domain. Critically, only the latter predicted individual choices, suggesting that the effect of stereotypes on behavior is mediated by inference-based decision-making processes in the OFC.

stereotypes | social perception | inequity aversion | fMRI | representational similarity analysis

Significance

People's decisions about how to treat others are known to be influenced by societally shared expectations about the typical traits of people from particular social groups (stereotypes). We combined a social psychological framework, an economic game, and multivariate functional MRI analysis to investigate whether and how trait inferences are instantiated neurally in the service of behavior toward members of different social groups. Multidimensional representations of trait content were found in brain regions associated with social cognition and in a region associated with inference-based decision-making: the lateral orbitofrontal cortex (OFC). Only OFC representations predicted individual participants' behavior, suggesting that although stereotypes are also represented in social cognition regions, they exert influence on behavior via decision-making mechanisms centered in the OFC.

In daily human life, people frequently make decisions about how to treat other people. Whether these decisions are fleeting (e.g., "Do I hold open the door for the approaching person?") or more consequential ("Whom should I hire?"), a hallmark of human social decision-making is flexibility: the ability to adapt our behavior to interactions with different individuals based on information about what those individuals are like. However, people's assumptions about what others are like are not always accurate. In particular, they are known to be influenced disproportionately by cues to the person's group membership (i.e., stereotypes), including the person's gender, age, race, nationality, or occupation (1–4), setting up the potential to perpetuate disparities in treatment across different social groups. Although an abundance of research in the behavioral sciences has examined when and how people stereotype others based on their group membership (5, 6) and documented treatment disparities in domains ranging from medicine to education (7), it has been a challenge to characterize the impact of stereotypes on social decision-making processes, including the computational mechanisms that mediate the influence of stereotype information on social behavior (8, 9).

A recent advance at the intersection of psychology and behavioral economics offers a new framework to test hypotheses about how stereotypes about others' traits are incorporated into neural computations that guide social behavior. This advance builds upon the observation that stereotypes are structured along core dimensions of trait perception, such as warmth (the degree to which people have good intentions toward others) and competence (the degree to which people are capable of acting on their intentions) (5, 6). We recently developed a model that characterizes how trait perceptions interact with the decision context to guide people's resource-allocation behavior (10) toward members of different social groups (8). Specifically, by incorporating stereotypes about others' warmth and competence into a computational model of social valuation (11, 12), we found that these two dimensions of stereotype content exerted dissociable effects on individuals' aversion to different forms of inequity: people were averse to receiving more money than were stereotypically warm others, and people were averse to receiving less money than were stereotypically competent others. In turn, this approach made it possible to predict with high accuracy not only individuals' behavior toward a wide variety of social groups in a laboratory setting but also people's treatment of members of different social groups in labor and education settings (8).

Author affiliations: [a]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; and [b]Haas School of Business and Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720

Author contributions: M.H. and A.C.J. designed research; A.C.J. performed research; K.K., J.W.K., and A.C.J. analyzed data; and K.K., J.W.K., M.H., and A.C.J. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: kenjik@sas.upenn.edu or acjenk@upenn.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2116944119/-/DCSupplemental.

Published May 23, 2022.

This evidence points to the possibility that assumptions about others' traits may be represented in the brain in a way that 1) corresponds to a dimensional structure of stereotype content and 2) enables stereotypes to exert influence on the computations underlying contextually flexible social decision-making. To test this, we used functional MRI (fMRI) and representational similarity analysis (RSA), along with a social decision task in which participants allocated resources across themselves and members of different social groups in decision contexts further characterized by advantageous or disadvantageous inequity.

Our hypotheses build upon our recent behavioral findings, along with previous neuroimaging research into trait perception and stereotyping, on the one hand, and into value-based decision-making, on the other. First, a consistent set of brain regions including the temporoparietal junction (TPJ), superior temporal sulcus (STS), and medial prefrontal cortex is activated when people think about the minds of others and is, therefore, sometimes referred to collectively as the *mentalizing network* (13–20). Activations of the mentalizing network have been observed across a wide range of social task paradigms, including those that require inference of others' traits based on their group membership (i.e., stereotyping) (21–25). However, it remains unclear whether and how these regions might mediate the effect of stereotypes on social decision-making, in large part because past studies of stereotyping have primarily involved passive viewing or basic judgments about others, making empirical characterization of behavior inapplicable; have focused mostly on how active different brain regions are, rather than on multidimensional trait representations (26); and have primarily involved judgments about a small number of social groups [e.g., males versus females (24, 27)], rather than a set of targets spanning the space of trait perception (28).

Second, value-based decision-making has long been associated with processes in a set of frontostriatal regions, including the ventral striatum, the ventromedial prefrontal cortex, and the orbitofrontal cortex (OFC) (28–33). A particularly intriguing area is the OFC, which is thought to guide decision-making by representing defining features of the task or environment, often not directly observable but inferred, that are critical for inferring or imagining future decision outcomes (29–36). Accordingly, the OFC may play a critical role in social behavior by representing inferences about others' traits in ways that are behaviorally relevant. If so, OFC processes could plausibly serve as a route through which trait representations inform inference-based evaluation of overall decision outcomes in social contexts, including how subjectively rewarding particular monetary allocations with particular recipients will be. This account has the potential to unify the seemingly independent effects observed in past studies of social decision-making, which have shown that choices in the laboratory and field are modulated by overt characteristics such as race (37), gender (38), and attractiveness (39, 40), by suggesting that they share a reliance on core, underlying representations of inferred trait content.

Here, we report evidence that neural representations of inferred trait content in the OFC systematically affect social decisions. Extending our previous behavioral findings (8), we find that recipients' inferred warmth (which we refer to here simply as "warmth") increases advantageous inequity aversion and their inferred competence (which we refer to here as "competence") increases disadvantageous inequity aversion. At the neural level, RSA revealed that stereotypic trait content was represented along the warmth and competence dimensions in the TPJ and STS, key regions in the mentalizing network, and in the OFC, a key region for goal-directed decision-making. Critically, we found that the representation in the OFC, but not in the other regions, predicted individual participants' context-dependent monetary-allocation decisions. This suggests that, while regions of the mentalizing network may be involved in inferences about others' traits, the effects of those trait perceptions on social decisions are mediated by inference-based decision-making mechanisms centered in the OFC.

Results

Experimental Paradigm. Participants ($n = 32$) played an extended version of the Dictator Game in an fMRI experiment. The participant played the role of Dictator and, in each trial, decided how to allocate money between themselves and a recipient. To experimentally manipulate the participant's inferences about the recipient's traits across trials, we provided one piece of information about the recipient's social group membership (e.g., their occupation or nationality). We selected 20 social groups to span a wide range of social perception along the trait dimensions of warmth and competence, and ratings of their warmth and competence were collected in an independent, online sample (Fig. 1*A*) (8). We also collected social perception ratings from our fMRI participants after scanning and confirmed they were highly consistent with the independent ratings (*SI Appendix*, Fig. S1), demonstrating the robustness of our social perception measures.

In each trial, the participant was presented with the information about the recipient (e.g., "Occupation: Nurse"; "Nationality: Japanese"), and then with two monetary-allocation options, between which they were asked to choose one (Fig. 1*B*). We manipulated these options so we could empirically characterize the tradeoff between decision-making motives (i.e., maximization of one's own payoff and concern for the inequity between oneself and the recipient). Specifically, in some trials, the participant chose between an equal allocation and an unequal allocation that created advantageous inequity (i.e., allocating more money to the participant than to the recipient); in other trials, the participant chose between an equal allocation and an unequal allocation that created disadvantageous inequity (i.e., allocating less money to the participant than to the recipient). This forced-choice design allowed us to directly examine how participants' preferences about advantageous and disadvantageous inequity depend on the recipient and, specifically, on inferences about the recipient's warmth and competence. To quantitatively measure the influence of these inferences on inequity aversion, the payoffs of the unequal allocation choice were manipulated across trials (three advantageous choices and three disadvantageous choices were presented for each recipient during scanning; see *Materials and Methods*).

Context-Dependent Effects of Others' Traits on Social Decisions. Behaviorally, the effects of recipients' warmth and competence on resource allocation depended on the decision context, such that warmth influenced choices in advantageous inequity trials, while competence influenced choices in disadvantageous inequity trials (Fig. 1*C*). In advantageous inequity trials, participants were less likely to choose the unequal allocation (and more likely to choose the equal allocation) when the recipient's warmth was higher (Pearson's $r = -0.60$; permutation $P = 0.004$). Their choices about advantageous inequity were not correlated with competence ($r = -0.09$; $P = 0.331$), and the effect of warmth was stronger than that of competence ($P = 0.004$). Conversely, in disadvantageous inequity trials,

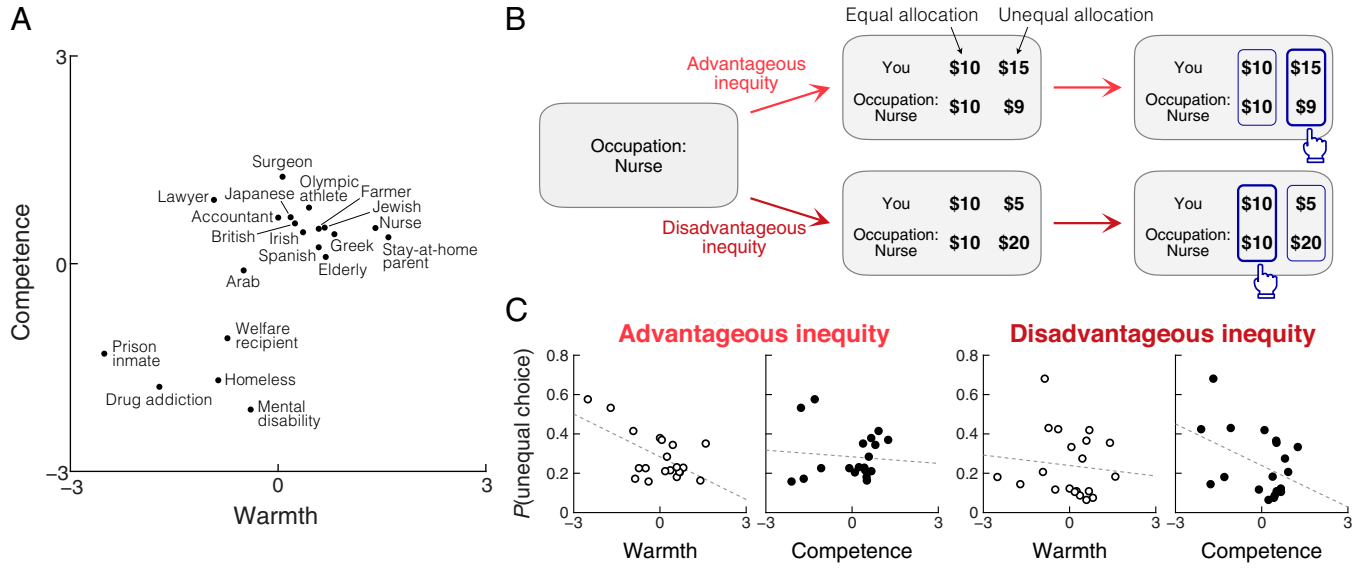

Fig. 1. Experimental paradigm and behavioral results. *A*. Recipients in the Dictator Game were identified by their social-group membership. A total of 20 social groups were chosen so that the recipient's warmth and competence were variable across trials. *B*. In each trial, the recipient's social group was first presented, followed by two allocation options, one equal and one unequal. The participant was asked to make a binary choice. The unequal option allocated more money to the participant than the recipient in advantageous inequity trials (*Top*) and less money in disadvantageous inequity trials (*Bottom*). *C*. Participants' allocation choices were influenced by the recipient's traits in a context-dependent manner. In advantageous inequity trials (*Left*), participants were less likely to choose the unequal option (and more likely to choose the equal option) when the recipient's warmth was higher ($r = -0.60$, permutation $P = 0.004$), irrespective of their competence ($r = -0.09$, $P = 0.331$). In disadvantageous inequity trials (*Right*), participants were less likely to choose the unequal option when the recipient's competence was higher ($r = -0.43$, $P = 0.040$), irrespective of their warmth ($r = -0.11$, $P = 0.307$).

participants were less likely to choose the unequal allocation when the recipient's competence was higher ($r = -0.43$; $P = 0.040$). Their choices about disadvantageous inequity were not correlated with warmth ($r = -0.11$; $P = 0.307$), and the effect of competence was stronger than that of warmth ($P = 0.049$). Therefore, aversion to advantageous inequity increased with the recipient's warmth, whereas aversion to disadvantageous inequity increased with the recipient's competence. These behavioral results replicate our previous findings (8) despite substantial differences in experimental design, including the use of binary forced choices between equal and unequal allocations (rather than continuous allocations) in the present study. Accordingly, they also correspond to our previously observed relationships of warmth and competence to behavior in field settings, pointing to the likely ecological validity of the present investigation.

Neural Representations of Others' Traits. Our behavioral findings show that inferences about other people's traits, guided by information about social groups and organized along distinct dimensions of warmth and competence, exert strong and dissociable effects on social decision-making processes as captured by our extended Dictator Game. Accordingly, we next looked for neural representations of these inferred traits. To elucidate the representation of traits and not payoff structures or decision processes, we focused on signals during the portion of each trial when the participant was presented with the recipient's group membership, prior to the presentation of the allocation options (Fig. 1*A*). In a representational similarity analysis (RSA), we looked for brain regions where two recipients who are assumed to have similar traits (e.g., an accountant and a Japanese person, who are both perceived to have high competence and moderate warmth) evoke similar response patterns, and two recipients who are assumed to have dissimilar traits (e.g., an accountant and a prison inmate) evoke dissimilar response patterns (41). We adopted a whole-brain searchlight approach that looked for brain regions where the representational dissimilarity matrix (RDM) of the local response patterns in a spherical searchlight was correlated with the RDM of the trait, defined by pairwise Euclidean distance in the two-dimensional space of warmth and competence (Fig. 2*A*). To construct the neural RDM, we quantified dissimilarity in response patterns, using cross-validated Mahalanobis distance, which is a metric of the extent to which response patterns evoked by different recipients are consistently distinguishable across scanning runs (42).

Our RSA revealed that recipients' warmth and competence are represented in the left lateral OFC, which has long been associated with inference-based decision-making (whole-brain family-wise error [FWE]–corrected threshold-free cluster enhancement [TFCE] $P < 0.05$). In addition to the OFC, traits are also represented in several other regions, including those associated with mentalizing, such as the right TPJ, left STS, left inferior frontal gyrus, left superior frontal gyrus, and right premotor cortex (Fig. 2*B* and *SI Appendix*, Table S1).

Linking Neural Trait Representations to Choice Behavior. Next, we investigated to what extent trait representations in these regions contributed to participants' subsequent monetary-allocation decisions (Fig. 3*A*). We reasoned that if representations in any of the trait-representing regions (Fig. 2*B*) contribute to decision-making, then individual variations in local neural responses in such a region should predict individual variation in allocation choices. More specifically, if two recipients evoke similar response patterns in a particular region of a particular participant's brain, and representations in that region contribute to decision-making in this context, then the participant should have treated those two recipients similarly. Likewise, recipients that evoke dissimilar response patterns in a given participant should have been treated dissimilarly by that participant. To test for such a relationship between neural responses and individual choices, we ran another RSA that examined the relationship between neural RDMs (on response patterns during the epoch of

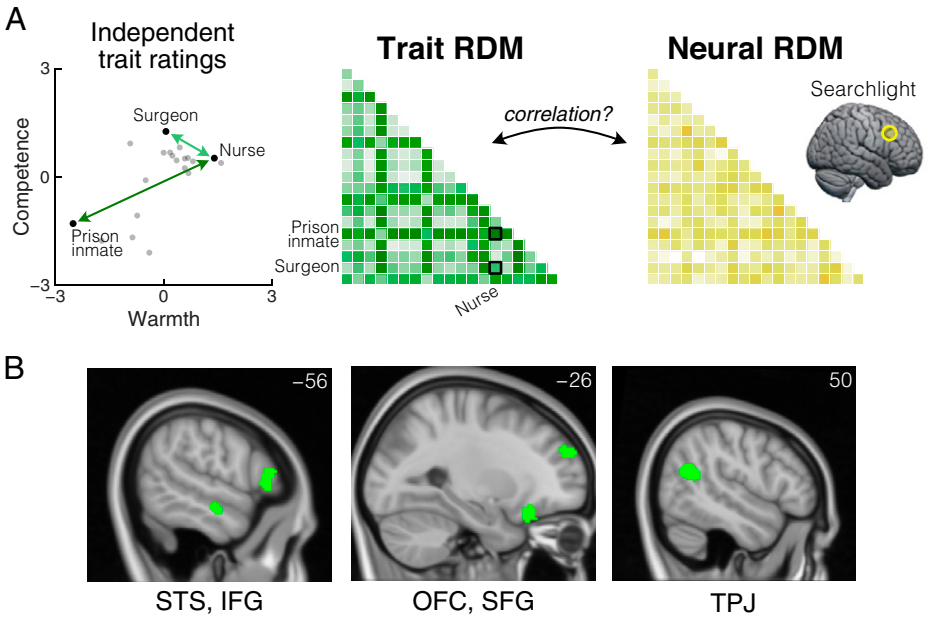

Fig. 2. Neural representations of others' traits. *A*. Whole-brain searchlight RSA looked for neural representations of the recipient's traits. The trait RDM was defined based on pairwise Euclidean distance in the two-dimensional space of warmth and competence. The neural RDM was computed for each searchlight based on pairwise cross-validated Mahalanobis distance between voxel-wise responses. *B*. Trait representation was found in left STS, left inferior frontal gyrus (IFG), left OFC, left superior frontal gyrus (SFG), and right TPJ (whole-brain FWE–corrected TFCE $P < 0.05$).

recipient identity presentation, as in the previous RSA) in each of the trait regions (Fig. 2*B*) and choice RDMs at the individual participant level (Fig. 3*A*). We visualized each participant's choice frequency against each recipient (i.e., in how many trials out of three they chose the unequal allocation over the equal allocation) in a two-dimensional space, with choices in advantageous inequity trials on one axis and choices in disadvantageous inequity on the other axis. Pairwise Euclidean distance in this choice space was used to construct the individual choice RDM. To test the correlation between individual choice RDMs and neural RDMs above and beyond the population-level effects of warmth and competence, we obtained a FWE–corrected null-hypothesis distribution via permutation (randomly pairing choice and neural RDMs from different participants).

This analysis revealed that only responses in the lateral OFC predicted individual allocation choices above what would be

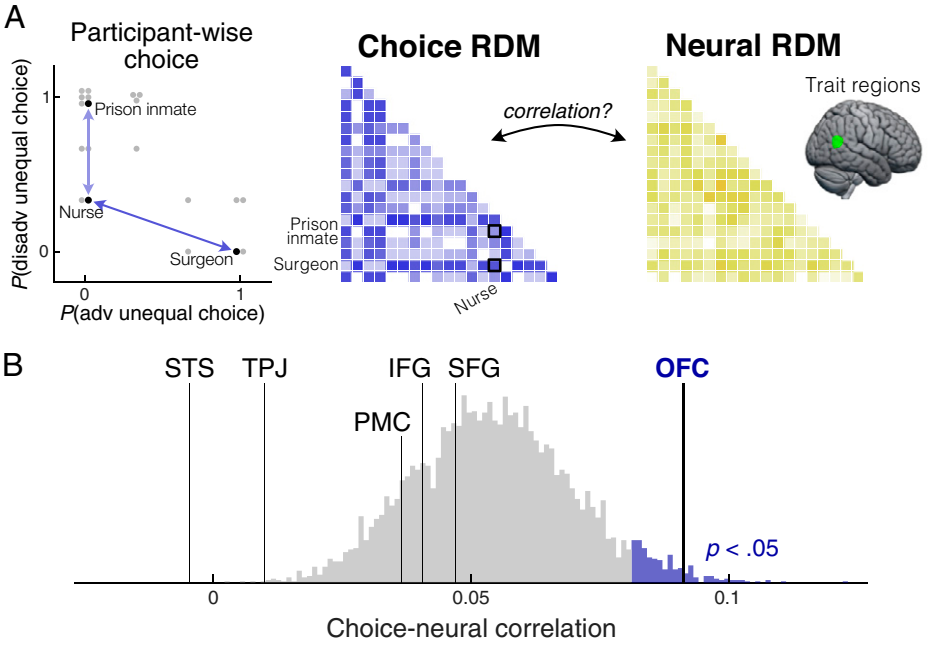

Fig. 3. Correlation between neural representations of traits and individual choices. (*A*) The relationship between individual-level allocation choices and response patterns in the regions that represent others' traits (identified in Fig. 2*B*) was evaluated in the second RSA. The choice RDM was constructed for each participant based on pairwise Euclidean distance in the two-dimensional space of choice frequency in advantageous and disadvantageous inequity trials. Its relationship with the neural RDM in each trait region was measured by *z*-transformed Spearman correlation. Shown are the data from one exemplar participant. *B*. The neural RDM in the OFC ($P = 0.011$), but not in any other region ($P > 0.50$), was significantly correlated with the individual-level choice RDM. Histogram: permutation-based FWE–corrected null hypothesis distribution. IFG, inferior frontal gyrus; PMC, premotor cortex; SFG, superior frontal gyrus.

expected by chance (for FWE-corrected across the regions of interest [ROIs], $P = 0.011$; Fig. 3*B*). No other region exhibited a significant relationship with choices ($P > 0.50$). This shows that the degree to which each participant treated recipients differently was correlated with the degree of differentiation in their representations in the OFC, suggesting that the trait representation in the lateral OFC contributes to the allocation decisions. Importantly, while our behavioral analysis revealed that the trait dimension (warmth or competence) that drives choices is dependent on the decision context (advantageous or disadvantageous inequity), responses in the lateral OFC were characterized by the two-dimensional spaces of traits (warmth and competence) and choices (advantageous and disadvantageous inequity) even before the participant was informed of the specific decision context. Taken together, these results suggest that the OFC plays a critical role in incorporating the perception of others' traits into social decision-making in a context-dependent manner.

Discussion

Adaptive social decision-making relies on information about others' traits and mental states. However, we often need to interact with people with whom we have very little experience. In such cases, people sometimes rely on inferences derived from societally shared stereotypes based on cues to others' social group membership (1–6, 8). Here, we identified a neural route through which stereotype content influences social decision-making. Using an extended Dictator Game paradigm in which participants allocated monetary resources between themselves and various recipients identified by information about their social group membership, we first showed that people spontaneously treated others differently depending on their traits in a way that hinged on aspects of the decision context; specifically, under advantageous inequity, choices depended on the recipient's warmth, while under disadvantageous inequity, choices depended on their competence. Using fMRI and RSA, we further showed that the recipients' traits were represented in brain regions associated with both mentalizing (the TPJ and STS) and inference-based decision-making (the OFC). Critically, the representation in the OFC was predictive of monetary-allocation choices at the individual level. Using a permutation test, we confirmed that this relationship cannot be accounted for by population-level effects of warmth and competence, and instead implies that individual differences in the OFC signals are associated with those in decision-making. This shows that the OFC plays an important role in driving social decisions based on the perception of others' traits.

Evidence that the lateral OFC mediates the effect of trait representations on social decision-making connects to a large body of evidence in humans and other species that the OFC contributes to decision-making that is guided by inference or imagination of outcomes. Previous studies used paradigms such as outcome devaluation or preconditioning to demonstrate that the OFC (in particular, the lateral OFC) is necessary for inference-based decision-making in rats (43, 44), monkeys (45, 46), and humans (47–49). Furthermore, recent neuroimaging and electrophysiological studies revealed that the OFC represents latent features of the environment, such as the hidden state of the present trial in sequential or learning tasks, that are not directly observable but are critical for outcome prediction (29, 30, 50–54). Based on this evidence, a current influential hypothesis posits that the OFC represents aspects of the environment that are not fully observable but are critical (or at least beneficial) for inference on future outcomes and thereby guides decision-making (31–35).

Our findings that the lateral OFC represents the traits of others, and that this representation is predictive of individual choices regarding these others, are consistent with the hypothesized function of the OFC. First, recipients' traits are not directly observable and instead inferred from information about their group membership. Second, and more important, inferences about recipients' traits affect inference-based evaluation of allocation outcomes, as demonstrated by the participants' revealed preference in the present study as well as our previous studies with independent samples (8). Taken together, these findings point to the possibility that the lateral OFC represents the recipient's traits in the present experimental paradigm because they are critical variables for inference-based evaluation of resource allocations; it is likely that the OFC does not represent others' traits in decision contexts that rely on other variables.

Other studies have also shown that the OFC is involved in incorporating perceptions of others' traits into social decisions in a task-dependent manner. For instance, racial features of faces are represented in the OFC when participants chose whether to befriend them but not when they judged whether they looked athletic (55), and patients with lateral OFC damage are able to judge competence of faces but fail to incorporate it into voting decisions (56). These findings, along with various social deficits exhibited by patients with OFC damage (35), show that the role of OFC in inference-based decision-making extends to the social domain. Indeed, inference-based outcome evaluation is critical for a wide range of social decisions, since the social world is characterized by a high degree of uncertainty with complex latent structures (e.g., who are friends and who are foes) and countless unobservable variables (e.g., beliefs and preferences of individuals) (57, 58).

We also found neural representations of recipients' traits in several regions outside the OFC. Among them, the right TPJ and the left STS are prominent areas in the mentalizing network, which is consistently activated when people infer others' traits, including based on their group membership (i.e., stereotyping) (21–25). Our results extend these previous findings by showing that multivoxel response patterns in the TPJ and STS contain multidimensional information about the traits of others. Interestingly, the STS (particularly its ventral bank, where we found trait representations) is anatomically connected to the lateral OFC in monkeys (59), raising the possibility that the behaviorally relevant representations in the OFC rely on inputs from the mentalizing network. In addition, the regions where we found trait representations outside the mentalizing network are also anatomically connected to the lateral OFC in monkeys (59–61), and many of these regions are also functionally coupled with the lateral OFC in resting-state and task-based fMRI in humans (62, 63). Taken together, these findings suggest that the use of stereotypes in social decision-making relies on interaction between two key systems: one anchored on the mentalizing network, which is responsible for inferences about others' traits, and the other primarily centered on the OFC, which incorporates the inferred traits flexibly into the evaluation of social-decision outcomes. This possibility is further supported by our finding that signals in the OFC, but not in other regions, are correlated with individual choices, which suggests that the OFC contributes to subsequent decision-making processes (64).

Our findings raise a number of exciting questions for future research. First, studies are needed to better understand the

circuit-level mechanisms through which multidimensional representations in the OFC drive subsequent decision-making processes. For example, it is possible that the context-specific effects of social perception on behavior (e.g., warmth affects advantageous inequity aversion, while competence affects disadvantageous inequity aversion) could be mediated by flexible readout of the OFC signals by downstream regions (65). Second, it remains an open question how trait representations in the mentalizing network and the OFC are constructed from semantic knowledge about social groups, possibly represented in the anterior temporal lobe (66–68). Third, while we did not find evidence of trait representations in the hippocampus, a previous study reported that self-other relationships are represented in the hippocampus in a two-dimensional ego-centric space (69). This raises the intriguing possibility that the OFC and hippocampus play complementary roles in social decision-making by representing the social world in different frames of reference (31, 32, 70–72). Finally, our findings have the potential to inform future inquiry into the neuroscience of discrimination, for example, by quantifying relationships between societal treatment of social groups and representations of their traits in the OFC (9, 73, 74), as well as into disorders of social function, for example, by separating social deficits arising from an atypical neural representation of others' traits from those arising from an atypical integration of trait representations into value-based decision-making (75).

Future research could also elucidate why trait representation was not observed in the medial prefrontal cortex in this study, at least at a standard statistical threshold for whole-brain analysis. Although the medial prefrontal cortex is also generally recruited during stereotyping (22–25) and mentalizing (15–19, 76, 77), it is possible that the MPFC contributes to stereotyping in a way that does not involve trait representations in a two-dimensional warmth–competence space (28, 72, 78, 79); that its contributions might be more specialized for inferences about individuals based on richer, more individuating information (80–83); or that its involvement depends on the degree to which mentalizing is explicitly called for. For example, previous studies reported that the medial prefrontal cortex is more activated when participants receive explicit instructions to mentalize (84), whereas the TPJ is consistently activated even when no explicit instructions or incentives for mentalizing are provided (76, 85, 86). These possibilities further highlight the potential importance of goals and incentives in understanding the neural basis of social decision-making.

More broadly, while the present study focused on stereotypes, this is not the only route to trait inference. For instance, people often assume that others tend to hold attitudes or beliefs like their own (social projection), particularly when making inferences about individuals who are perceived to be similar to themselves (4, 18, 82, 83, 87). Furthermore, for individuals with whom people interact extensively, trait information can be accumulated by learning from experience (66, 88, 89). It remains an open question how trait information acquired through these different routes affects social decisions at the cognitive and neural levels. For its part, the present study establishes how stereotypes drive social decisions via task-relevant representations in the OFC, forming the basis for a more comprehensive understanding of the neural mechanisms through which different types of social inferences affect social decisions across different contexts.

Materials and Methods

All procedures were approved by the institutional review boards at the University of California, Berkeley, and Virginia Tech.

Participants. A total of 43 healthy people provided informed consent in accordance with the Declaration of Helsinki and participated in the experiment. Data from one participant were removed because of image artifacts, and data from an additional 10 participants were removed because of excessive motion (showing frame-wise or cumulative displacement of >2 mm in translation or >2.5 degrees in rotation), leaving data from 32 participants for analysis (22 women and 10 men, 18–64 years old, mean age [SD] = 27.5 [11.4]).

Task Overview. Participants chose how to allocate monetary resources between themselves and a series of recipients in a modified Dictator Game. In each trial, the participant viewed one piece of social-group information about the recipient for that trial (e.g., nurse, Japanese), along with two allocation options. In a majority of trials, one of the options provided an equal division of resources between the participant and the recipient, while the other option provided an unequal division of resources favoring either the participant (advantageous inequity) or the recipient (disadvantageous inequity). In the remaining trials, both options provided equal divisions in different amounts; these trials were only included to encourage the participant to pay attention to both sets of payoffs and were not included in the primary analyses in this study (*SI Appendix*, Fig. S2 *C* and *D* for behavioral data in these trials). In all cases, the participant decided unilaterally which option to choose, while the recipient had no ability to affect the outcome.

Recipient Identities. The recipient was described by one of 20 social-group memberships, which were originally developed in our previous study (8) to span a wide range of trait perceptions along the core dimensions of warmth and competence. The group membership was described by one of the following attributes: occupation (accountant, surgeon, lawyer, nurse, stay-at-home parent, Olympic athlete, farmer), nationality (Japanese, Irish, British, Spanish, Greek), ethnicity (Jewish, Arab), medical history (mental disability), age demographic (elderly), psychiatric history (drug addiction), housing status (homeless), financial status (welfare recipient), and legal status (prison inmate). The group membership was presented along with the attribute (e.g., "Occupation: Nurse" or "Nationality: Japanese").

In all behavioral and fMRI analyses, we used ratings of these recipients' warmth and competence collected from an independent sample in an online experiment [$n = 252$; study 1b in our previous study (8)]. To confirm that this independently measured social perception was shared by participants in the current fMRI experiment, we also asked these participants to rate recipients' warmth and competence after the scan. We confirmed that the average ratings obtained in the present study were highly correlated with the independent ratings, demonstrating the robustness of our social perception measures (*SI Appendix*, Fig. S1).

Monetary-Allocation Options. While the equal allocation option provided the same amount of money to the participant and the recipient ($10) across all trials, payoffs in the unequal allocation option varied across trials. The payoff structure was either ($20, $5), ($15, $9), or ($14, $6) in advantageous inequity trials and either ($5, $20), ($9, $15), or ($6, $14) in disadvantageous inequity trials, where the first number indicates the own payoff and the second number indicates the recipient's payoff. Therefore, in the advantageous inequity trials, the participant could maximize their own payoff by choosing the unequal allocation and maximize the recipient's payoff by choosing the equal allocation. Conversely, in the disadvantageous inequity trials, they could maximize their own payoff by choosing the equal allocation and maximize the recipient's payoff by choosing the unequal allocation. We varied the payoff structure orthogonally with the recipient manipulation across trials to reliably detect differences in inequity aversion across recipients, based on previously documented ranges of inequity aversion parameters and their individual differences (90–92).

Procedure. Participants completed the task inside the MRI scanner and indicated their choices using a button box. The task was programmed in Python using the *Pygame* package. Prior to scanning, participants were instructed that although the monetary allocations in this task were hypothetical, they should indicate as honestly as possible which choice they would prefer if it were to affect the actual payoffs of themselves and the recipient. Throughout scanning, each of eight payoff structures was presented once for each of the 20 recipients. In total, $8 \times 20 = 160$ trials were presented in a randomized order for each participant.

The scanning consisted of two runs (80 trials each), with each recipient appearing four times per run.

In each trial, the participant was first presented with the recipient information only (duration of presentation was drawn randomly from a uniform distribution from 2.5 s to 5.5 s), and then with two allocation options, presented side by side. To mitigate cognitive load, the constant equal allocation (i.e., $10 to the recipient and to oneself) was always presented to the left, while the right option was varied across trials. After a delay (varied between 3 s and 6 s), both options were outlined by blue boxes, which prompted the participant to indicate a choice by pressing one of two buttons. Participants were asked to press a button within 5 s; the trial was automatically terminated (and not repeated) when they did not press a button within that time window.

Behavioral Data Analysis. Economic theories of distributional preference posit that decision-making in the Dictator Game is driven primarily by two factors: maximization of one's own payoff and concern for the inequity between one's own payoff and the recipient's payoff (11, 12). They further posit that preferences regarding advantageous inequity are distinct from preferences regarding disadvantageous inequity (90, 91). In recent work, we found that aversion to advantageous inequity increases with the recipient's warmth (but does not depend on their competence) and aversion to disadvantageous inequity increases with the recipient's competence (but does not depend on their warmth) (8). In that study, the participant decided how many tokens to share with the recipient in a continuous manner, and thus it was up to them whether and how often they created advantageous or disadvantageous inequity. We adopted a different task design in the present study, which used two-alternative forced choices regarding advantageous and disadvantageous inequity in separate trials. This design allowed us to test the dissociable effects of warmth and competence on inequity preference even more directly.

We counted how often the participants chose the unequal allocation over the equal allocation against each recipient in advantageous and disadvantageous inequity trials and tested their correlation with the warmth and competence of the recipients for those choices (Fig. 1*C*). The statistical significance of the correlation was assessed via permutation (9,999 iterations). The same permutation test was also used to assess whether the effects of warmth and competence on choice frequencies were different from each other (i.e., whether the difference in coefficients between the behavior–warmth correlation and the behavior–competence correlation was statistically significant compared with the null-hypothesis distribution). While Fig. 1*C* shows choice frequencies marginalized over payoff structures in each trial type, the relationship with trait perceptions was robustly observed even when measured for each payoff structure separately (*SI Appendix*, Fig. S2 *A* and *B*).

MRI Data Acquisition. Magnetic resonance images were acquired by a 3T Siemens Magnetom Trio scanner and a 12-channel head coil. A three-dimensional, high-resolution structural image was acquired using a T1-weighted, magnetization-prepared, rapid-acquisition gradient-echo pulse sequence (voxel size = 1 × 1 × 1 mm; matrix size = 190 × 239; 200 axial slices; repetition time = 2,300 msec; echo time = 2.98 msec). While participants completed the task, functional images were acquired using a T2*-weighted gradient echo-planar imaging pulse sequence (voxel size = 3 × 3 × 3 mm; interslice gap = 0.15 mm; matrix size = 64 × 64; 32 oblique axial slices; repetition time = 2,000 msec; echo time = 30 msec). Slices were angled +30° with respect to the anterior commissure–posterior commissure line to reduce signal dropout in the OFC (93).

MRI data analysis: trait perception. We conducted a whole-brain searchlight (RSA to look for neural representations of the recipient's traits (41). More specifically, we looked for brain regions in which voxel-wise local response patterns evoked by two recipients were similar (or dissimilar) when inferences about their traits were also similar (or dissimilar) to each other. Our RSA formulated this relationship as the correlation between two RDMs: one that captures dissimilarity in trait perception (trait RDM) and one that captures dissimilarity in response patterns (neural RDM), in all possible pairs of recipients (20 recipients; 190 pairwise similarity measures).

For the trait RDM, pairwise dissimilarity in traits was quantified as Euclidean distance in a two-dimensional space of warmth and competence (Fig. 1*A*). Empirical measures of warmth and competence perceptions were originally obtained as numeric scores between 0 and 100 (8). We used *z*-scores computed across the 20 recipients for each dimension to construct the Euclidean space.

The neural RDM was computed at every voxel within gray matter in native space. Pairwise dissimilarity in voxel-wise response patterns was quantified as the cross-validated Mahalanobis (Crossnobis) distance in a gray-matter spherical searchlight (10-mm radius). Crossnobis distance is an unbiased measure of the extent to which response patterns evoked by two recipients are consistently distinguishable across scanning runs (42). We chose this distance measure rather than alternative measures because we were primarily interested in how recipients are distinguished in their neural representation, rather than how they are similarly represented. In our experiment, since each recipient was presented four times in each of the two scanning runs, we were able to cross-validate distance estimates across runs to mitigate spurious distance caused by noise.

The pairwise Crossnobis distance was estimated following the formulae provided previously (42). We first estimated voxel-wise response patterns evoked by each recipient in each scanning run using a generalized linear model (GLM) implemented in SPM12 software. To retain fine-grained signals as much as possible, minimal preprocessing (specifically, only motion correction) was applied to echo-planar images prior to the GLM. The GLM included the regressors of interest, modeling the presentation of each recipient using a boxcar function that starts with the onset of the recipient presentation and ends with the onset of payoffs presentation, along with nuisance regressors modeling button presses. These regressors were convolved with the canonical double-gamma hemodynamic response function and its temporal derivative. The GLM also included confound regressors for head motion (3 translations and 3 rotations, estimated in the motion-correction procedure), 128-s high-pass filtering, and a first-order serial autoregression model [AR(1)]. The GLM coefficients of each recipient within the searchlight were then cross-validated across the two runs to obtain the Crossnobis distance. For Mahalanobis whitening, we estimated the covariance matrix in the searchlight using the GLM residuals and shrank it for invertibility (94).

We computed Fisher-transformed Spearman correlation between the trait and neural RDMs at each gray-matter voxel. We discovered that the trait RDM inadvertently contained information about visual features of the recipient presentation on the screen and, specifically, its character count. This visual confound was controlled by partialling out another RDM that captured the character count. The resultant correlation map was normalized to the standard Montreal Neurological Institute (MNI) space based on the magnetization-prepared, rapid-acquisition gradient-echo structural image of each participant and was spatially smoothed (Gaussian kernel full width at half maximum = 8 mm) using SPM12 software. For the population-level analysis, a cluster-level permutation test was conducted using the FSL randomise tool (whole-brain FWE- corrected TFCE $P < 0.05$; 4,999 iterations).

MRI data analysis: correlation with individual choices. To look for evidence that any of the regions that represented the recipients' traits (Fig. 2*B* and *SI Appendix*, Table S1) contributed to the subsequent monetary-allocation decisions, we ran another RSA, which tested the correlation between neural RDMs and choice RDMs. We predicted that, if a region contributed to the decisions, local response patterns evoked by two recipients in one participant's brain would be similar (or dissimilar) to each other when the participant treated them in a similar (or dissimilar) manner in their allocation choices.

The individual-choice RDM was built on the frequency with which each participant chose the advantageous or disadvantageous unequal allocation for each recipient. Pairwise Euclidean distance was measured in the two-dimensional space of the observed choice frequencies, one dimension for advantageous inequity trials and the other dimension for disadvantageous inequity trials. Since each recipient was presented in three advantageous inequity trials and three disadvantageous inequity trials, the choice frequency on each dimension was either 0, 1/3, 2/3, or 1.

These individual-level choice RDMs were then correlated with neural RDMs in the regions identified by our first RSA as containing representations of others' traits. Binary masks were functionally defined in standard MNI space based on the aforementioned population-level statistics (whole-brain FWE-corrected TFCE $P < 0.05$) and converted to the native space of each participant's brain using SPM12 software. The *z*-transformed Spearman correlation between the choice and neural RDMs was averaged across all voxels in the native-space masks.

To test whether neural response patterns predicted individual choice patterns above and beyond the population-level effects of warmth and competence, we conducted a permutation test, randomly pairing choice and neural RDMs from different participants (4,999 iterations). To control for multiple comparisons across ROIs, the nullhypothesis distribution was constructed by taking the highest population average of correlation scores across the ROIs in each permutation iteration.

Data Availability. Anonymized, raw data and analysis code are deposited in OSF (https://osf.io/qkrhm/). Raw fMRI data are available from Openneuro at https://openneuro.org/datasets/ds004128/.

ACKNOWLEDGMENTS We thank Nakyung Lee and Pierre Karashchuk for assistance with paradigm development; Duy Phan, Amanda Savarese, and Cassandra Carrin for assistance with data collection; and Dilara Berkay for helpful input and assistance with the preparation of fMRI data for analysis.

1. A. G. Greenwald, M. R. Banaji, Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **102**, 4–27 (1995).
2. S. E. Asch, Forming impressions of personality. *J. Abnorm. Psychol.* **41**, 258–290 (1946).
3. A. G. Greenwald, C. K. Lai, Implicit social cognition. *Annu. Rev. Psychol.* **71**, 419–445 (2020).
4. D. R. Ames, Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *J. Pers. Soc. Psychol.* **87**, 340–353 (2004).
5. A. E. Abele, N. Ellemers, S. T. Fiske, A. Koch, V. Yzerbyt, Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychol. Rev.* **128**, 290–314 (2021).
6. S. T. Fiske, A. J. C. Cuddy, P. Glick, Universal dimensions of social cognition: Warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
7. M. Bertrand, E. Duflo, "Field experiments on discrimination" in *Handbook of Field Experiments*, A. Banerjee, E. Duflo, Eds. (Elsevier, 2017), pp. 309–394.
8. A. C. Jenkins, P. Karashchuk, L. Zhu, M. Hsu, Predicting human behavior toward members of different social groups. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9696–9701 (2018).
9. D. M. Amodio, The neuroscience of prejudice and stereotyping. *Nat. Rev. Neurosci.* **15**, 670–682 (2014).
10. J. Andreoni, J. Miller, Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**, 737–753 (2002).
11. G. Charness, M. Rabin, Understanding social preferences with simple tests. *Q. J. Econ.* **117**, 817–869 (2002).
12. E. Fehr, K. M. Schmidt, A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868 (1999).
13. M. Schurz, J. Radua, M. Aichhorn, F. Richlan, J. Perner, Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* **42**, 9–34 (2014).
14. R. Saxe, N. Kanwisher, People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage* **19**, 1835–1842 (2003).
15. D. M. Amodio, C. D. Frith, Meeting of minds: The medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* **7**, 268–277 (2006).
16. R. N. Spreng, R. A. Mar, A. S. N. Kim, The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *J. Cogn. Neurosci.* **21**, 489–510 (2009).
17. C. D. Frith, U. Frith, Interacting minds–A biological basis. *Science* **286**, 1692–1695 (1999).
18. A. C. Jenkins, J. P. Mitchell, "How has cognitive neuroscience contributed to social psychological theory?" in *Social Neuroscience: Towards Understanding the Underpinnings of the Social Mind*, A. Todorov, S. Fiske, D. Prentice, Eds. (Oxford University Press, 2011), pp. 3–13.
19. P. Molenberghs, H. Johnson, J. D. Henry, J. B. Mattingley, Understanding the minds of others: A neuroimaging meta-analysis. *Neurosci. Biobehav. Rev.* **65**, 276–291 (2016).
20. R. B. Mars *et al.*, On the relationship between the "default mode network" and the "social brain". *Front. Hum. Neurosci.* **6**, 189 (2012).
21. J. M. Contreras, M. R. Banaji, J. P. Mitchell, Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Soc. Cogn. Affect. Neurosci.* **7**, 764–770 (2012).
22. L. Van der Cruyssen, E. Heleven, N. Ma, M. Vandekerckhove, F. Van Overwalle, Distinct neural correlates of social categories and personality traits. *Neuroimage* **104**, 336–346 (2015).
23. J. M. Contreras, J. Schirmer, M. R. Banaji, J. P. Mitchell, Common brain regions with distinct patterns of neural responses during mentalizing about groups and individuals. *J. Cogn. Neurosci.* **25**, 1406–1417 (2013).
24. S. Quadflieg *et al.*, Exploring the neural correlates of social stereotyping. *J. Cogn. Neurosci.* **21**, 1560–1570 (2009).
25. J. Delplanque, E. Heleven, F. Van Overwalle, Neural representations of groups and stereotypes using fMRI repetition suppression. *Sci. Rep.* **9**, 3190 (2019).
26. D. I. Tamir, M. A. Thornton, J. M. Contreras, J. P. Mitchell, Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 194–199 (2016).
27. J. P. Mitchell, D. L. Ames, A. C. Jenkins, M. R. Banaji, Neural correlates of stereotype application. *J. Cogn. Neurosci.* **21**, 594–604 (2009).
28. L. T. Harris, S. T. Fiske, Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychol. Sci.* **17**, 847–853 (2006).
29. N. W. Schuck, M. B. Cai, R. C. Wilson, Y. Niv, Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016).
30. R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
31. C. Padoa-Schioppa, K. E. Conen, Orbitofrontal cortex: A neural circuit for economic decisions. *Neuron* **96**, 736–754 (2017).
32. A. M. Wikenheiser, G. Schoenbaum, Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. *Nat. Rev. Neurosci.* **17**, 513–523 (2016).
33. T. A. Stalnaker, N. K. Cooch, G. Schoenbaum, What the orbitofrontal cortex does not do. *Nat. Neurosci.* **18**, 620–627 (2015).
34. Y. Niv, Learning task-state representations. *Nat. Neurosci.* **22**, 1544–1553 (2019).
35. L. Q. Yu, I. P. Kan, J. W. Kable, Beyond a rod through the skull: A systematic review of lesion studies of the human ventromedial frontal lobe. *Cogn. Neuropsychol.* **37**, 97–141 (2020).
36. M. C. Klein-Flügge, H. C. Barron, K. H. Brodersen, R. J. Dolan, T. E. J. Behrens, Segregated encoding of reward-identity and stimulus-reward associations in human orbitofrontal cortex. *J. Neurosci.* **33**, 3202–3211 (2013).
37. J. F. Dovidio, K. Kawakami, C. Johnson, B. Johnson, A. Howard, On the nature of prejudice: Automatic and controlled processes. *J. Exp. Soc. Psychol.* **33**, 510–540 (1997).
38. A. Ben-Ner, B. P. McCall, M. Stephane, H. Wang, Identity and in-group/out-group differentiation in work and giving behaviors: Experimental evidence. *J. Econ. Behav. Organ.* **72**, 153–170 (2009).
39. D. S. Hamermesh, J. E. Biddle, Beauty and the labor market. *Am. Econ. Rev.* **84**, 1174–1194 (1994).
40. M. M. Mobius, T. S. Rosenblat, Why beauty matters. *Am. Econ. Rev.* **96**, 222–235 (2006).
41. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis - Connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
42. A. Walther *et al.*, Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* **137**, 188–200 (2016).
43. M. Gallagher, R. W. McMahan, G. Schoenbaum, Orbitofrontal cortex and representation of incentive value in associative learning. *J. Neurosci.* **19**, 6610–6614 (1999).
44. J. L. Jones *et al.*, Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953–956 (2012).
45. A. Izquierdo, R. K. Suda, E. A. Murray, Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *J. Neurosci.* **24**, 7540–7548 (2004).
46. E. A. West, J. T. DesJardin, K. Gale, L. Malkova, Transient inactivation of orbitofrontal cortex blocks reinforcer devaluation in macaques. *J. Neurosci.* **31**, 15128–15135 (2011).
47. J. Reber *et al.*, Selective impairment of goal-directed decision-making following lesions to the human ventromedial prefrontal cortex. *Brain* **140**, 1743–1756 (2017).
48. F. Wang, J. D. Howard, J. L. Voss, G. Schoenbaum, T. Kahnt, Targeted stimulation of an orbitofrontal network disrupts decisions based on inferred, not experienced outcomes. *J. Neurosci.* **40**, 8726–8733 (2020).
49. J. D. Howard *et al.*, Targeted stimulation of human orbitofrontal networks disrupts outcome-guided behavior. *Curr. Biol.* **30**, 490–498.e4 (2020).
50. S. C. Y. Chan, Y. Niv, K. A. Norman, A probability distribution over latent causes, in the orbitofrontal cortex. *J. Neurosci.* **36**, 7817–7828 (2016).
51. J. Zhou *et al.*, Evolving schema representations in orbitofrontal ensembles during learning. *Nature* **590**, 606–611 (2021).
52. T. A. Stalnaker, N. Raheja, G. Schoenbaum, Orbitofrontal state representations are related to choice adaptations and reward predictions. *J. Neurosci.* **41**, 1941–1951 (2021).
53. M. R. Nassar, J. T. McGuire, H. Ritz, J. W. Kable, Dissociable forms of uncertainty-driven representational change across the human brain. *J. Neurosci.* **39**, 1688–1698 (2019).
54. I. Saez *et al.*, Encoding of multiple reward-related computations in transient and sustained high-frequency activity in human OFC. *Curr. Biol.* **28**, 2889–2899.e3 (2018).
55. S. J. Gilbert, J. K. Swencionis, D. M. Amodio, Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia* **50**, 3600–3611 (2012).
56. C. Xia, D. Stolle, E. Gidengil, L. K. Fellows, Lateral orbitofrontal cortex links social impressions to political choices. *J. Neurosci.* **35**, 8507–8514 (2015).
57. A. C. Jenkins, J. P. Mitchell, Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cereb. Cortex* **20**, 404–410 (2010).
58. U. R. Karmarkar, A. C. Jenkins, "Neural and behavioral insights into online trust and uncertainty" in *Organizational Neuroethics*, J. Martineau and E. Racine (Springer, 2020), pp. 191–207.
59. S. T. Carmichael, J. L. Price, Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. *J. Comp. Neurol.* **363**, 642–664 (1995).
60. K. S. Saleem, H. Kondo, J. L. Price, Complementary circuits connecting the orbital and medial prefrontal networks with the temporal, insular, and opercular cortex in the macaque monkey. *J. Comp. Neurol.* **506**, 659–693 (2008).
61. S. T. Carmichael, J. L. Price, Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *J. Comp. Neurol.* **371**, 179–207 (1996).
62. D. H. Zald *et al.*, Meta-analytic connectivity modeling reveals differential functional connectivity of the medial and lateral orbitofrontal cortex. *Cereb. Cortex* **24**, 232–248 (2014).
63. T. Kahnt, L. J. Chang, S. Q. Park, J. Heinzle, J.-D. Haynes, Connectivity-based parcellation of the human orbitofrontal cortex. *J. Neurosci.* **32**, 6240–6250 (2012).
64. S. Ballesta, W. Shi, K. E. Conen, C. Padoa-Schioppa, Values encoded in orbitofrontal cortex are causally related to economic choices. *Nature* **588**, 450–453 (2020).
65. D. Birman, J. L. Gardner, A flexible readout mechanism of human sensory representations. *Nat. Commun.* **10**, 3500 (2019).
66. D. M. Amodio, Social cognition 2.0: An interactive memory systems account. *Trends Cogn. Sci.* **23**, 21–33 (2019).
67. D. M. Amodio, M. Cikara, The social neuroscience of prejudice. *Annu. Rev. Psychol.* **72**, 439–469 (2021).
68. I. R. Olson, D. McCoy, E. Klobusicky, L. A. Ross, Social cognition and the anterior temporal lobes: A review and theoretical framework. *Soc. Cogn. Affect. Neurosci.* **8**, 123–133 (2013).
69. R. M. Tavares *et al.*, A map for social navigation in the human brain. *Neuron* **87**, 231–243 (2015).
70. T. E. J. Behrens *et al.*, What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
71. S. A. Park, D. S. Miller, H. Nili, C. Ranganath, E. D. Boorman, Map making: Constructing, combining, and inferring on abstract cognitive maps. *Neuron* **107**, 1226–1238.e8 (2020).
72. S. A. Park, D. S. Miller, E. D. Boorman, Inferences on a multidimensional social hierarchy use a grid-like code. *Nat. Neurosci.* **24**, 1292–1301 (2021).
73. J. T. Kubota, M. R. Banaji, E. A. Phelps, The neuroscience of race. *Nat. Neurosci.* **15**, 940–948 (2012).
74. B. D. Mattan, K. Y. Wei, J. Cloutier, J. T. Kubota, The social neuroscience of race-based and status-based prejudice. *Curr. Opin. Psychol.* **24**, 27–34 (2018).

75. K. Gray, A. C. Jenkins, A. S. Heberlein, D. M. Wegner, Distortions of mind perception in psychopathology. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 477–479 (2011).
76. F. Van Overwalle, M. Vandekerckhove, Implicit and explicit social mentalizing: Dual processes driven by a shared neural network. *Front. Hum. Neurosci.* **7**, 560 (2013).
77. M. Schurz *et al.*, Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychol. Bull.* **147**, 293–327 (2021).
78. F. Van Overwalle, N. Ma, K. Baetens, Nice or nerdy? The neural representation of social and competence traits. *Soc. Neurosci.* **11**, 567–578 (2016).
79. M. Li *et al.*, Warmth is more influential than competence: An fMRI repetition suppression study. *Brain Imaging Behav.* **15**, 266–275 (2021).
80. F. Van Overwalle, Social cognition and the brain: A meta-analysis. *Hum. Brain Mapp.* **30**, 829–858 (2009).
81. E. Heleven, F. Van Overwalle, The person within: Memory codes for persons and traits using fMRI repetition suppression. *Soc. Cogn. Affect. Neurosci.* **11**, 159–171 (2016).
82. J. P. Mitchell, C. N. Macrae, M. R. Banaji, Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron* **50**, 655–663 (2006).
83. A. C. Jenkins, C. N. Macrae, J. P. Mitchell, Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4507–4512 (2008).
84. J. P. Mitchell, M. R. Banaji, C. N. Macrae, The link between social cognition and self-referential thought in the medial prefrontal cortex. *J. Cogn. Neurosci.* **17**, 1306–1315 (2005).
85. S. Boccadoro *et al.*, Defining the neural correlates of spontaneous theory of mind (ToM): An fMRI multi-study investigation. *Neuroimage* **203**, 116193 (2019).
86. J. Kestemont, M. Vandekerckhove, N. Ma, N. Van Hoeck, F. Van Overwalle, Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Soc. Cogn. Affect. Neurosci.* **8**, 481–493 (2013).
87. D. I. Tamir, J. P. Mitchell, Anchoring and adjustment during social inferences. *J. Exp. Psychol. Gen.* **142**, 151–162 (2013).
88. L. M. Hackel, B. B. Doll, D. M. Amodio, Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nat. Neurosci.* **18**, 1233–1235 (2015).
89. P. Mende-Siedlecki, Y. Cai, A. Todorov, The neural dynamics of updating person impressions. *Soc. Cogn. Affect. Neurosci.* **8**, 623–631 (2013).
90. Y. Morishima, D. Schunk, A. Bruhin, C. C. Ruff, E. Fehr, Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron* **75**, 73–79 (2012).
91. A. Bruhin, E. Fehr, D. Schunk, The many faces of human sociality: Uncovering the distribution and stability of social preferences. *J. Eur. Econ. Assoc.* **72**, 738 (2018).
92. X. Gao *et al.*, Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7680–E7689 (2018).
93. N. Weiskopf, C. Hutton, O. Josephs, R. Deichmann, Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at 3 T and 1.5 T. *Neuroimage* **33**, 493–504 (2006).
94. O. Ledoit, M. Wolf, Honey, I shrunk the sample covariance matrix. *J. Portfol. Manage.* **30**, 110–119 (2004).