

Neural representations of stereotype content predict social decisions

Kenji Kobayashi¹, Joseph W. Kable¹, Ming Hsu², & Adrianna C. Jenkins¹

¹University of Pennsylvania

²University of California, Berkeley

Address for correspondence:

Adrianna C. Jenkins or Kenji Kobayashi

acjenk@upenn.edu

kenjik@sas.upenn.edu

1 **Abstract**

2
3 Perceptions of others' traits based on social group membership (stereotypes) are known
4 to affect social behavior, but little is known about the neural mechanisms mediating
5 these effects. Here, using fMRI and representational similarity analysis (RSA), we
6 investigated neural representations of others' traits and their contributions to social
7 decision making. Behaviorally, perceptions of others' traits, captured by a two-
8 dimensional framework, biased participants' monetary allocation choices in a context-
9 dependent manner: recipients' perceived warmth increased advantageous inequity
10 aversion and competence increased disadvantageous inequity aversion. Neurally, RSA
11 revealed that stereotypes about others' traits were represented in activity patterns in the
12 temporoparietal junction and superior temporal sulcus, two regions associated with
13 mentalizing, and in the lateral orbitofrontal cortex (OFC), known to represent latent
14 environmental features during goal-directed outcome inference. Critically, only the latter
15 predicted individual choices, suggesting that the effect of stereotypes on behavior is
16 mediated by inference-based, goal-directed decision-making processes in the OFC.

17 **Introduction**

18
19 People approach their interactions with different individuals using information
20 about those individuals' traits. For example, people might be generous toward a friend
21 who is generally appreciative but cautious toward a coworker who is generally
22 untrustworthy. However, people often interact with individuals with whom they lack
23 extensive experience, meaning they often need to make inferences about others based
24 on indirect information. One common route to such inferences is to rely on societally
25 shared perceptions of people's traits associated with their social group membership,
26 such as their nationality or occupation (i.e., stereotypes)¹⁻⁴. Such inferences can
27 provide a shortcut to what would otherwise be highly uncertain social decision making,
28 but widespread reliance on stereotypes is also associated with societal treatment
29 disparities⁵. Although an abundance of research in the behavioral sciences has
30 examined when and how people stereotype others based on their group membership⁶,
31 only recently have we begun to understand the behavioral consequences of stereotypes
32 and the computational mechanisms mediating these effects⁷. Even less is known about
33 neural representations of others' traits and their contributions to social decision making.

34 Mounting evidence from social psychology shows that stereotypes are structured
35 along core dimensions of trait perception, such as warmth, or the degree to which
36 people have good intentions toward others, and competence, or the degree to which
37 people are capable of acting on their intentions^{6,8}. Recently, we adopted a novel
38 modeling approach that enabled us to characterize empirically how these trait
39 perceptions contribute to people's decisions about how to treat others⁷. In a modified

40 Dictator game, people made monetary allocation decisions between themselves and
41 other individuals (recipients) from various social groups^{9–11}. By incorporating people’s
42 perceptions of recipients’ warmth and competence into a computational model of social
43 valuation, we discovered that these dimensions of social perception exerted dissociable,
44 context-dependent effects on individuals’ aversion to different forms of inequity. When
45 paired with recipients perceived as more warm, people were more strongly averse to
46 advantageous inequity (i.e., receiving more than the recipient), whereas when paired
47 with recipients perceived as more competent, they were more strongly averse to
48 disadvantageous inequity (i.e., receiving less than the recipient). Furthermore, this
49 approach was able to quantitatively predict the complex pattern of disparities observed
50 in field experiments in labor markets and education settings.

51 These findings raise the possibilities that the human brain represents stereotypes
52 about others’ traits along core dimensions, including warmth and competence, and that
53 these representations systematically bias social decisions in a context-dependent
54 manner. Despite a wealth of neuroimaging research on trait perception and
55 stereotyping, on the one hand, and on value-based decision making, on the other, these
56 possibilities remain empirically untested. In particular, past studies of social perception
57 have primarily involved passive viewing or basic judgments in non-decision contexts,
58 making empirical characterization of behavior inapplicable. Additionally, such studies
59 have focused mostly on *how active* different brain regions are, rather than on multi-
60 dimensional trait representations¹², and have primarily involved judgments about a small
61 number of social groups (e.g., males versus females), rather than a set of targets

62 spanning the space of social perception^{13,14}. Likewise, although past studies of social
63 decision-making have shed light on how choice processes are modulated by overt
64 characteristics such as race, gender, and attractiveness, it remains unclear to what
65 extent these effects are related, and much is unknown about the underlying
66 mechanisms.

67 Here we investigate the neural mechanisms underlying the effect of stereotypes
68 about others traits on social decisions using fMRI and representational similarity
69 analysis (RSA). A brain region that mediates the effect of perceived traits on behavior
70 should, at a minimum, represent traits in some way, such that those generally perceived
71 to have more similar traits produce more similar response patterns. However, not all
72 regions that represent traits as such must necessarily play a role in the translation of
73 trait information into behavior. We predict that, if trait representations in some brain
74 regions are linked directly to behavior, idiosyncratic similarity in response patterns in
75 those regions can be used to predict variations in context-dependent choices across
76 individuals.

77 Past neuroscientific studies suggest two, non-mutually exclusive hypotheses.
78 One possibility is that a set of regions widely associated with social cognition, such as
79 the temporoparietal junction (TPJ), superior temporal sulcus (STS), and medial
80 prefrontal cortex (MPFC), represent others' traits in the service of social behavior.
81 These regions are consistently activated when people attempt to infer the mental states
82 of others and are therefore often referred to collectively as the mentalizing network^{15–22}.
83 Activations of the mentalizing network have been observed across a wide range of task

84 paradigms, including those that require inference of others' traits based on their group
85 membership (i.e., stereotyping)^{13,23–26}. Although it has not yet been tested directly, these
86 mentalizing regions may contribute to social decision making by representing others'
87 perceived traits along core dimensions, including warmth and competence²⁷.

88 Another possibility is that perceived traits of others are represented in
89 frontostriatal regions involved in social and non-social value-based decision making,
90 such as the ventral striatum, the ventromedial prefrontal cortex, and the orbitofrontal
91 cortex (OFC)^{28–33}. In particular, the OFC is thought to guide flexible, goal-directed
92 decisions by representing defining features of the task or environment, often not directly
93 observable but inferred, that are critical for inferring or imagining future decision
94 outcomes^{34–40}. Given that stereotyping plays a particular role when people interact with
95 people with whom we do not have extensive experience, the OFC may contribute to
96 social decisions by representing their inferred traits and thereby enabling inference-
97 based evaluation of decision outcomes (e.g., how subjectively rewarding particular
98 monetary allocations with particular recipients will be). Therefore, frontostriatal regions
99 involved in decision making, and in particular the OFC, may play a critical role in social
100 behavior by representing others' traits when they are behaviorally relevant.

101 To test these possibilities, we conducted an fMRI experiment using an adapted
102 version of our previous paradigm to investigate neural representations of other's traits in
103 the service of monetary allocation decisions. We show that, consistent with our previous
104 behavioral finding⁷, recipient's perceived warmth increases advantageous inequity
105 aversion and perceived competence increases disadvantageous inequity aversion. At

106 the neural level, RSA revealed that traits were represented along the warmth and
107 competence dimensions in the TPJ and STS, key regions in the mentalizing network,
108 and in the OFC, a key region for goal-directed decision making. Critically, we
109 discovered that the representation in the OFC, but not in the other regions, predicted
110 individual participants' monetary allocation decisions. This suggests that, while the
111 mentalizing network may be involved in inferences about others' traits, the effects of
112 those trait perceptions on social decisions are mediated by domain-general
113 mechanisms of inference-based, goal-directed decision making centered in the OFC.

114

115 **Results**

116 **Experimental paradigm**

117 Participants ($n = 32$) played an extended version of the Dictator game in an fMRI
118 experiment. The participant played the role of Dictator and, on each trial, decided how
119 to allocate money between themselves and a recipient. To experimentally manipulate
120 the participant's perception of the recipient's traits across trials, we provided one piece
121 of information about the recipient's social group membership (e.g., their occupation or
122 nationality). We selected 20 social groups to span a wide range of social perception
123 along the trait dimensions of warmth and competence, and ratings of their warmth and
124 competence were collected in an independent, online sample⁷. We also collected social
125 perception ratings from our fMRI participants after scanning and confirmed that they
126 were highly consistent with the independent ratings (Fig. S1), demonstrating the
127 robustness of our social perception measures.

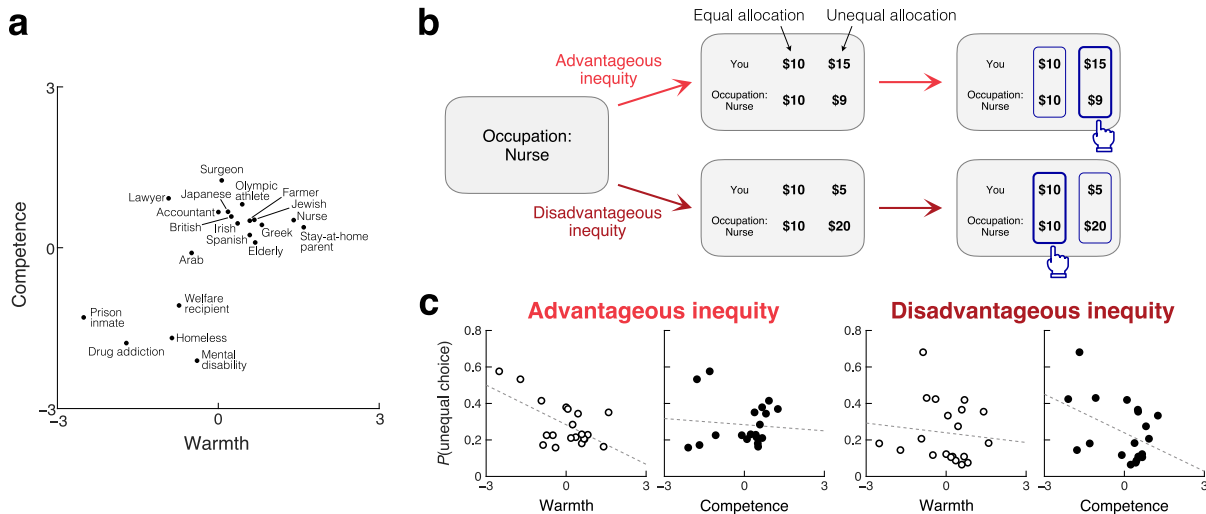


Fig 1. Experimental paradigm and behavioral results. **a.** Recipients in the Dictator game were identified by their social group membership. 20 social groups were chosen so that the recipient's perceived warmth and competence were variable across trials. **b.** On each trial, the recipient's social 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 inequality trials (*top*) and less money in disadvantageous inequality trials (*bottom*). **c.** Participants' allocation choices were influenced by the recipient's perceived traits in a context-dependent manner. *Left:* In advantageous inequality trials, participants were less likely to choose the unequal option (and more likely to choose the equal option) when the recipient's perceived warmth was higher ($r = -.60$, permutation $p = .004$), irrespective of their competence ($r = -.09$, $p = .331$). *Right:* in disadvantageous inequality trials, participants were less likely to choose the unequal option when the recipient's perceived competence was higher ($r = -.43$, $p = .040$), irrespective of their warmth ($r = -.11$, $p = .307$).

128 On each trial, the participant was presented with the information about the

129 recipient (e.g., "Occupation: Nurse"; "Nationality: Japanese"), and then with two

130 monetary allocation options, between which they were asked to choose one (Fig. 1b).

131 We manipulated these options so that we could empirically characterize the tradeoff

132 between decision-making motives, i.e., maximization of one's own payoff and concern

133 for the inequity between oneself and the recipient. Specifically, in some trials, the

134 participant chose between an equal allocation and an unequal allocation that created

135 advantageous inequality (i.e., allocating more money to the participant than to the

136 recipient); in other trials, the participant chose between an equal allocation and an

137 unequal allocation that created disadvantageous inequality (i.e., allocating less money to

138 the participant than to the recipient). This forced choice design allowed us to directly
139 examine how participants' preferences about advantageous and disadvantageous
140 inequity depend on the recipient, and specifically, on the recipient's perceived warmth
141 and competence.

142

143 **Context-dependent effects of others' traits on social decisions**

144 Behaviorally, the recipients' perceived warmth and competence exerted diverging
145 effects on participants' monetary allocation decisions; perceived warmth influenced
146 choices in advantageous inequity trials, while perceived competence influenced choices
147 in disadvantageous inequity trials (Fig. 1c). In advantageous inequity trials, participants
148 were less likely to choose the unequal allocation (and more likely to choose the equal
149 allocation) when the recipient's perceived warmth was higher (Pearson's $r = -.60$,
150 permutation $p = .004$). Their choices about advantageous inequity were not correlated
151 with perceived competence ($r = -.09$, $p = .331$), and the effect of warmth was stronger
152 than that of competence ($p = .004$). Conversely, in disadvantageous inequity trials,
153 participants were less likely to choose the unequal allocation when the recipient's
154 perceived competence was higher ($r = -.43$, $p = .040$). Their choices about
155 disadvantageous inequity were not correlated with perceived warmth ($r = -.11$, p
156 $= .307$), and the effect of competence was stronger than that of warmth ($p = .049$).
157 Therefore, aversion to advantageous inequity increases with the recipient's warmth,
158 whereas aversion to disadvantageous inequity increases with the recipient's
159 competence. These behavioral results replicate our previous findings⁷ despite

160 substantial differences in experimental design, including the use of binary forced
161 choices between equal and unequal allocations (rather than continuous allocations) in
162 the current study.

163

164 **Neural representations of others' traits**

165 Our behavioral findings show that perceptions of other people's traits, guided by
166 information about social groups and organized along distinct dimensions of warmth and
167 competence, exert strong and dissociable effects on social decision-making processes
168 as captured by our extended Dictator game. Accordingly, we next looked for neural
169 representations of these perceived traits. To elucidate the representation of perceived
170 traits and not payoff structures or decision processes, we focused on BOLD signals
171 during the portion of each trial when the participant was presented with the recipient's
172 group membership, prior to the presentation of the allocation options (Fig. 1a). We
173 looked for brain regions where two recipients that are similar to each other in perceived
174 traits (e.g., an Accountant and a Japanese person, who are both perceived to have high
175 competence and moderate warmth) evoke similar response patterns, and two recipient
176 that are dissimilar in perceived traits (e.g., an Accountant and a Prison inmate) evoke
177 dissimilar response patterns (representational similarity analysis; RSA⁴¹). We adopted a
178 whole-brain searchlight approach that looked for brain regions where the
179 representational dissimilarity matrix (RDM) of the local response patterns in a spherical
180 searchlight was correlated with RDM of the perceived trait, defined by pairwise
181 Euclidean distance in the two-dimensional space of warmth and competence (Fig. 2a).

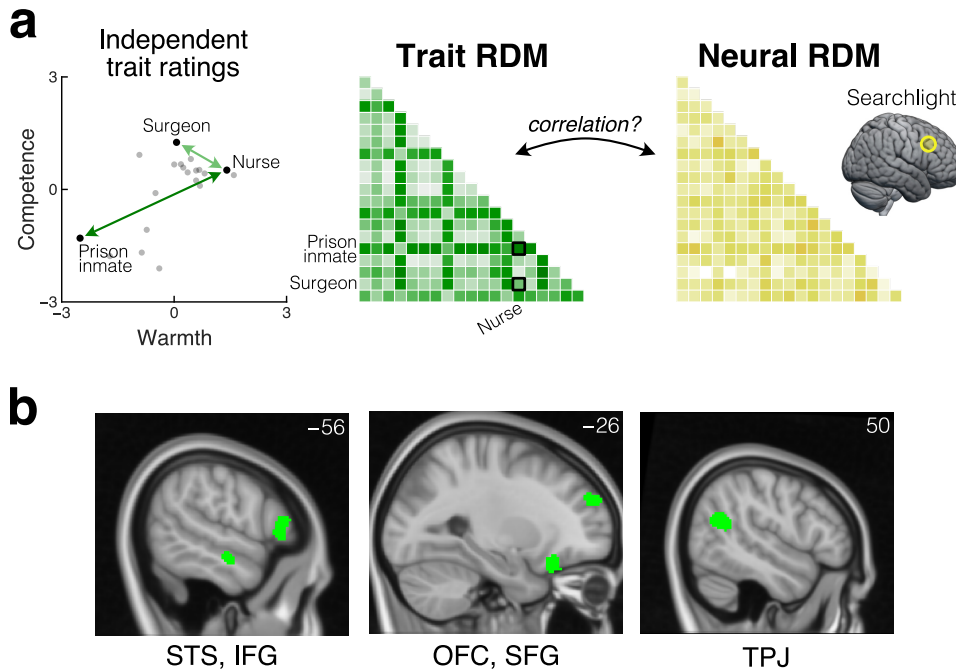


Fig 2. Neural representations of others' traits. **a.** Whole-brain searchlight RSA looked for neural representations of the recipient's perceived traits. The trait RDM was defined based on pairwise Euclidean distance in the two-dimensional space of warmth and competence. The neural RDM was computed for each searchlight based on pairwise cross-validated Mahalanobis distance between voxel-wise responses. **b.** Trait representation was found in left STS, left IFG, left OFG, left SFG, right TPJ, and right PMC (not shown) (whole-brain FWE-corrected TFCE $p < .05$).

182 To construct the neural RDM, we quantified dissimilarity in response patterns using
183 cross-validated Mahalanobis distance, which is a metric of the extent to which response
184 patterns evoked by different recipients are consistently distinguishable across scanning
185 runs⁴².

186 Our RSA revealed that recipients' perceived warmth and competence are
187 represented in left lateral orbitofrontal cortex (OFC), which has long been associated
188 with inference-based, goal-directed decision making (threshold-free cluster
189 enhancement [TFCE], whole-brain family-wise error [FWE] corrected $p < .05$). In
190 addition to the OFC, perceived traits are also represented in several other regions,
191 including those associated with mentalizing, such as the right temporoparietal junction

192 (TPJ), left superior temporal sulcus (STS), left inferior frontal gyrus, left superior frontal
193 gyrus, and right premotor cortex (Fig. 2b).

194

195 **Linking neural trait representations to choice behavior**

196 Next, we investigated to what extent trait representations in these regions
197 contributed to participants' subsequent monetary allocation decisions (Fig. 3a). We
198 reasoned that, if representations in any of the trait-representing regions (Fig. 2b)
199 contribute to decision making, then individual variations in local neural responses in
200 such a region should predict individual variation in allocation choices. More specifically,

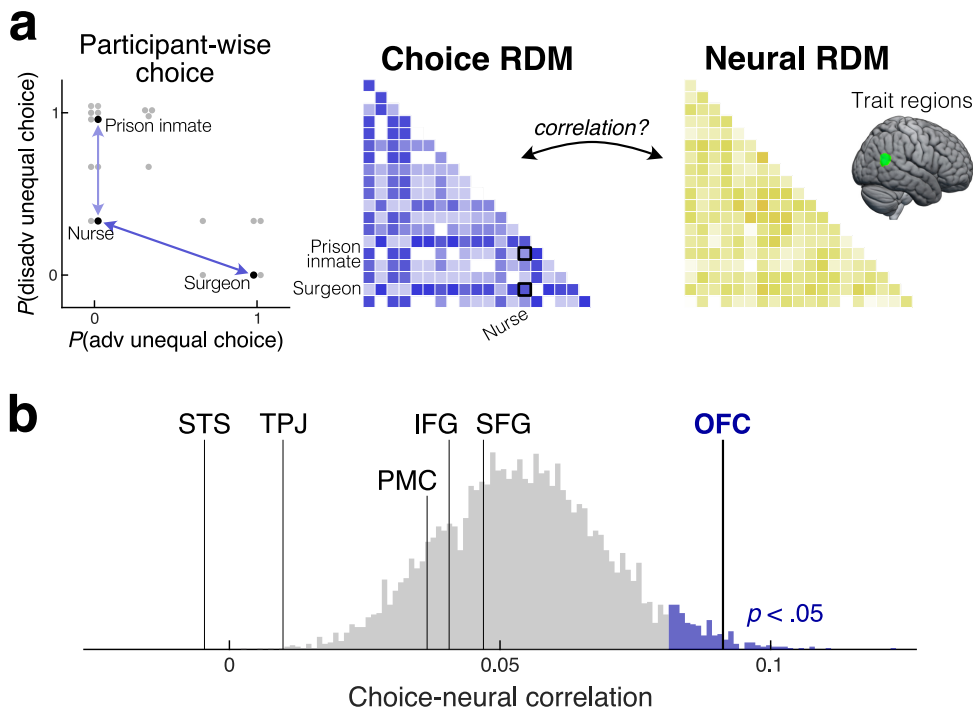


Fig 3. Correlation between neural representations of traits and individual choices. a.

Relationship between individual-level allocation choices and response patterns in the regions that represent others' traits (identified in Fig. 2b) 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 is the data from one exemplar participant. b. The neural RDM in the OFC ($p = .011$), but not in any other region ($p > .50$), was significantly correlated with the individual-level choice RDM. Histogram: permutation-based FWE-corrected null hypothesis distribution.

201 if two recipients evoke similar response patterns in a particular region of a particular
202 participant's brain, and representations in that region contribute to decision-making in
203 this context, then the participant should have treated those two recipients similarly.
204 Likewise, recipients that evoke dissimilar response patterns in a given participant should
205 have been treated dissimilarly by that participant. To test for such a relationship
206 between neural responses and individual choices, we ran another RSA that examined
207 the relationship between neural RDMs (on response patterns during the epoch of
208 recipient identity presentation, as in the previous RSA) in each of the trait regions (Fig.
209 2b) and choice RDMs at the individual subject level (Fig. 3a). We visualized each
210 participant's choice frequency against each recipient (i.e., how often they chose the
211 unequal allocation over the equal allocation) as a two-dimensional space, with choices
212 in advantageous inequity trials on one axis and choices in disadvantageous inequity on
213 the other axis. Pairwise Euclidean distance in this choice space was used to construct
214 the individual choice RDM. To test the correlation between individual choice RDMs and
215 neural RDMs above and beyond the population-level effects of warmth and
216 competence, we obtained an FWE-corrected null-hypothesis distribution via permutation
217 (randomly pairing choice and neural RDMs from different participants).

218 This analysis revealed that only responses in the lateral OFC predicted individual
219 allocation choices above chance (FWE corrected across the ROIs, $p = .011$; Fig. 3b).
220 No other region exhibited a significant relationship with choices ($p > .50$). This suggests
221 that the representation of the recipient's traits in the lateral OFC contributes to the
222 allocation decisions. Importantly, while our behavioral analysis revealed that the trait

223 dimension (warmth or competence) that drives choices is *dependent on the decision*
224 *context* (advantageous or disadvantageous inequity), responses in the lateral OFC were
225 characterized by the two-dimensional spaces of traits (warmth and competence) and
226 choices (advantageous and disadvantageous inequity), even before the participant was
227 informed of the specific decision context. Taken together, these results suggest that the
228 OFC plays a critical role in incorporating the perception of others' traits into social
229 decision making in a highly flexible, goal-directed, context-dependent manner.

230

231 **Discussion**

232 Adaptive social decision making relies on inferences about others' traits and
233 mental states. However, we often need to interact with people with whom we have very
234 little experience. In such cases, people sometimes rely on societally shared
235 stereotypes, or trait perceptions based on cues to social group membership^{1-4,6-8}. Here
236 we identified a neural mechanism through which such trait perceptions influence social
237 decision making. Using an extended Dictator game paradigm in which participants
238 allocated monetary resources between themselves and various recipients identified by
239 information about their social group membership, we first showed that people
240 spontaneously treat others differently depending on their perceived traits in a context-
241 dependent manner; advantageous inequity aversion increased with the recipient's
242 warmth, while disadvantageous inequity aversion increased with their competence.
243 Using fMRI and RSA, we further showed that the recipients' traits were represented in
244 brain regions associated with both mentalizing (TPJ and STS) and goal-directed

245 decision making (OFC). Critically, the representation in the OFC was predictive of
246 monetary allocation choices at the individual level. Using a permutation test, we
247 confirmed that this relationship cannot be accounted for by population-level effects of
248 warmth and competence, and instead implies that individual differences in the OFC
249 signals are associated with those in decision making. This shows that the OFC plays an
250 important role in driving social decisions based on the perception of others' traits.

251 Evidence that the lateral OFC mediates the effect of trait representations on
252 social decision-making connects to a large body of evidence in humans and other
253 species that the OFC contributes to goal-directed behavior. Goal-directed behavior is
254 guided by inferred or imagined outcomes, as opposed to habitual behavior that is
255 guided by cached values learned through trial and error. Previous studies used
256 paradigms such as outcome devaluation or preconditioning to demonstrate that the
257 OFC (in particular the lateral OFC) is necessary for goal-directed behavior in rats^{43,44},
258 monkeys^{45,46}, and humans⁴⁷⁻⁴⁹. Furthermore, recent neuroimaging and
259 electrophysiological studies revealed that the OFC represents latent features of the
260 environment, such as the hidden state of the current trial in sequential or learning tasks,
261 that are not directly observable but are critical for outcome prediction^{34,35,50-54}. Based on
262 this evidence, a current influential hypothesis posits that the OFC represents aspects of
263 the environment that are not fully observable but critical (or at least beneficial) for
264 inference on future outcomes, and thereby guides flexible, goal-directed decision
265 making³⁶⁻⁴⁰.

266 Our findings, that the lateral OFC represents the perceived traits of others, and
267 that this representation is predictive of individual choices regarding these others, are
268 consistent with the hypothesized function of the OFC. First, recipients' traits are not
269 directly observable and instead inferred from information about their group membership.
270 Second, decisions in the current paradigm are guided by inferences about how
271 subjectively rewarding it would be to allocate money between the self and the recipient,
272 as opposed to trial-and-error learning. Third, and most important, perceived traits affect
273 inference-based evaluation of allocation outcomes, as demonstrated by the participants'
274 revealed preference in the current study as well as our previous studies with
275 independent samples⁷. Taken together, this points to the possibility that the lateral OFC
276 represent the recipient's traits in the current experimental paradigm because they are
277 critical variables for inference-based evaluation of resource allocations; it is likely that
278 the OFC does not represent others' traits in decision contexts that rely on other
279 variables.

280 Other studies have also shown that the OFC is involved in incorporating
281 perceptions of others' traits into social decisions in a goal-directed manner. For
282 instance, racial features of faces are represented in the OFC when participants chose
283 whether to befriend them (goal-directed decision making) but not when they judged
284 whether they looked athletic (not goal-directed decision making)⁵⁵, and patients with
285 lateral OFC damage are able to judge competence of faces but fail to incorporate it into
286 voting decisions⁵⁶. These findings, along with various social deficits exhibited by
287 patients with OFC damage⁴⁰, show that the role of OFC in inference-based, goal-

288 directed decision making extends to the social domain. Indeed, inference-based
289 outcome evaluation is critical for a wide range of social decisions, since the social world
290 is characterized by a high degree of uncertainty with complex latent structures (e.g.,
291 who are friends and who are foes) and countless unobservable variables (e.g., beliefs
292 and preferences of individuals)^{57,58}.

293 We also found neural representations of recipients' traits in several regions
294 outside the OFC. Among them, the right TPJ and the left STS are prominent areas in
295 the mentalizing network, which is consistently activated when people infer others' traits,
296 including based on their group membership (i.e., stereotyping)^{13,23–26}. Our results extend
297 these previous findings by showing, for the first time to our knowledge, that multi-voxel
298 response patterns in the TPJ and STS contain multi-dimensional information about
299 perceived traits of others. Interestingly, the STS (particularly its ventral bank, where we
300 found trait representations) is anatomically connected to the lateral OFC in monkeys⁵⁹,
301 raising the possibility that the goal-directed representations in the OFC rely on inputs
302 from the mentalizing network. In addition, the regions where we found trait
303 representations outside the mentalizing network are also anatomically connected to the
304 lateral OFC in monkeys^{59–61}, and many of these regions are also functionally coupled
305 with the lateral OFC in resting-state and task-based fMRI in humans^{62,63}. Taken
306 together, these findings suggest that the use of stereotypes in social decision making
307 relies on interaction between two key systems: one anchored on the mentalizing
308 network, which is responsible for inferences about others' traits, and the other primarily
309 centered on the OFC, which incorporates the inferred traits into outcome inferences and

310 evaluation in a context-dependent, goal-directed manner. This account is further
311 supported by our finding that signals in the OFC, but not in other regions, are correlated
312 with individual choices, which suggests that the OFC contributes to subsequent
313 decision-making processes⁶⁴.

314 Our findings open up a number of exciting questions for future research. First,
315 future studies are needed to better understand the circuit-level mechanisms through
316 which multi-dimensional representations in the OFC drive subsequent decision-making
317 processes. For example, it is possible that the context-specific effects of social
318 perception on behavior (warmth affects advantageous inequity aversion, while
319 competence affects disadvantageous inequity aversion) could be mediated by flexible
320 readout of the OFC signals by downstream regions⁶⁵. Second, it remains an open
321 question how trait representations in the mentalizing network and the OFC are
322 constructed from semantic knowledge about social groups, possibly represented in the
323 anterior temporal lobe^{66–68}. Third, while we did not find evidence of trait representations
324 in the hippocampus, a previous study reported that self-other relationships in a two-
325 dimensional ego-centric space is represented in the hippocampus⁶⁹. This raises the
326 intriguing possibility that the OFC and hippocampus play complementary roles in social
327 decision making by representing the social world in different frames of
328 reference^{36,37,70,71}. Finally, our findings have the potential to inform future inquiry into the
329 neuroscience of discrimination, for example by quantifying relationships between
330 societal treatment of social groups and representations of their traits in the OFC^{72–74}, as
331 well as into disorders of social function, for example by separating social deficits arising

332 from an atypical neural representation of others' traits from those arising from an
333 atypical integration of trait representations into value-based decision-making⁷⁵.

334 Future research could also elucidate why trait representation was not observed in
335 the MPFC in this context, at least at a standard statistical threshold for whole-brain
336 analysis. Although the MPFC is also generally recruited during stereotyping^{13,24–26} and
337 mentalizing^{17–21,76,77}, it is possible that the MPFC contributes to stereotyping in a way
338 that does not involve trait representations in a two-dimensional warmth-competence
339 space^{27,78,79}; that its contributions might be more specialized for inferences about
340 individuals based on richer, more individuating information^{80–83}; or that its involvement
341 depends on the degree to which mentalizing is explicitly called for. For example,
342 previous studies reported that the MPFC is more activated when participants receive
343 explicit instructions to mentalize⁸⁴, whereas the TPJ is consistently activated even when
344 no explicit instructions or incentives for mentalizing are provided^{76,85,86}. These
345 possibilities further highlight the potential importance of goals and incentives in
346 understanding the neural basis of social decision-making.

347 More broadly, while the current study focused on stereotypes, this is not the only
348 route to trait inference. For instance, people often assume that others tend to hold
349 attitudes or beliefs like their own (social projection), particularly when making inferences
350 about individuals that are perceived to be similar to themselves^{4,20,82,83,87}. Furthermore,
351 for individuals with whom people interact extensively, trait information can be
352 accumulated across learning from experience^{66,88,89}. It remains an open question how
353 trait information acquired through these different routes impacts social decisions at the

354 cognitive and neural levels. For its part, the current study establishes how stereotypes
355 drive social decisions via goal-directed representations in the OFC, forming the basis for
356 a more comprehensive understanding of the neural mechanisms through which different
357 types of social inferences affect social decisions across different contexts.

358 **Materials and Methods**

359
360 All procedures were approved by the Institutional Review Boards at the University of
361 California, Berkeley, and Virginia Tech.

362
363 **Participants** 43 healthy people provided informed consent in accordance with the
364 Declaration of Helsinki and participated in the experiment. Data from 1 participant were
365 removed for image artifacts and data from an additional 10 participants were removed
366 for excessive motion (showing frame-wise or cumulative displacement of >2mm in
367 translation or >2.5 degrees in rotation), leaving data from 32 participants for analysis (22
368 female, 10 male, age: 18-64, mean = 27.5, standard deviation = 11.4).

369
370 **Task overview** Participants chose how to allocate monetary resources between
371 themselves and a series of recipients in a modified dictator game. On each trial, the
372 participant viewed one piece of social group information about the recipient for that trial
373 (e.g., nurse, Japanese), along with two allocation options. In a majority of trials, one of
374 the options provided an equal division of resources between the participant and the
375 recipient, while the other option provided an unequal division of resources favoring
376 either the participant (advantageous inequity) or the recipient (disadvantageous
377 inequity). In the remaining trials, both options provided equal divisions in different
378 amounts; these trials were only included to encourage the participant to pay attention to
379 both sets of payoffs and were not included in the primary analyses in this paper (see
380 Fig. S2c, d for behavioral data in these trials). In all cases, the participant decided
381 unilaterally which option to choose, while the recipient had no ability to affect the
382 outcome.

383
384 **Recipient identities** The recipient was described by one of 20 social group
385 memberships, which were originally developed in our previous study⁷ to span a wide
386 range of trait perceptions along the core dimensions of warmth and competence. The
387 group membership was described by one of the following attributes: occupation
388 (accountant, surgeon, lawyer, nurse, stay-at-home parent, Olympic athlete, farmer),
389 nationality (Japanese, Irish, British, Spanish, Greek), ethnicity (Jewish, Arab), medical
390 history (mental disability), age demographic (elderly), psychiatric history (drug
391 addiction), housing status (homeless), financial status (welfare recipient), and legal
392 status (prison inmate). The group membership was presented along with the attribute,
393 e.g., "Occupation: Nurse" or "Nationality: Japanese".

394
395 In all behavioral and fMRI analyses, we used ratings of these recipients' warmth and
396 competence collected from an independent sample in an online experiment ($n = 252$,
397 Study 1b in our previous study⁷). To confirm that this independently measured social
398 perception was shared by participants in the current fMRI experiment, we also asked
399 these participants to rate recipients' warmth and competence after the scan. We
400 confirmed that the average ratings obtained in the current study were highly correlated

401 with the independent ratings, demonstrating the robustness of our social perception
402 measures (Fig. S1).

403

404 **Monetary allocation options** While the equal allocation option provided the same
405 amount to the participant and the recipient (\$10) across all trials, payoffs in the unequal
406 allocation option were varied across trials. The payoff structure ([own payoff, the
407 recipient's payoff]) was either [\$20, \$5], [\$15, \$9], or [\$14, \$6] in advantageous inequity
408 trials, and either [\$5, \$20], [\$9, \$15], or [\$6, \$14] in disadvantageous inequity trials.
409 Therefore, in the advantageous inequity trials, the participant can maximize their own
410 payoff by choosing the unequal allocation and maximize the recipient's payoff by
411 choosing the equal allocation. Conversely, in the disadvantageous inequity trials, they
412 can maximize their own payoff by choosing the equal allocation and maximize the
413 recipient's payoff by choosing the unequal allocation.

414

415 **Procedure** Participants completed the task inside the MRI scanner and indicated their
416 choices using a button box. The task was programmed in python using the Pygame
417 package. Prior to scanning, participants were instructed that, although the monetary
418 allocations in this task were hypothetical, they should indicate as honestly as possible
419 which choice they would prefer if it were to affect the actual payoffs of themselves and
420 the recipient. Throughout scanning, each of 8 payoff structures was presented once for
421 each of the 20 recipients; in total, $8 \times 20 = 160$ trials were presented in a randomized
422 order for each participant. The scanning consisted of two runs (80 trials each), with
423 each recipient appearing four times per run.

424

425 In each trial, the participant was first presented with the recipient information (duration
426 between 2.5 sec to 5.5 sec: varied across scanning runs and participants), and then
427 with two allocation options, presented side by side. To mitigate cognitive load, the
428 constant equal allocation [\$10, \$10] was always presented to the left, while the right
429 option was varied across trials. After a delay (jittered between 3 sec and 6 sec), both
430 options were outlined by blue boxes, which prompted the participant to indicate a choice
431 by pressing one of two buttons. Participants were asked to press a button within 5
432 seconds; the trial was automatically terminated (and not repeated) when they did not
433 press a button within that window.

434

435 **Behavioral data analysis** Economic theories of distributional preference posit that
436 decision making in the Dictator game is driven primarily by two factors: maximization of
437 one's own payoff and concern for the inequity between one's own payoff and the
438 recipient's payoff^{10,11}. They further posit that preferences regarding advantageous
439 inequity are distinct from preferences regarding disadvantageous inequity^{90,91}. In recent
440 work, we found that aversion to advantageous inequity increases with the recipient's
441 perceived warmth (but does not depend on their perceived competence) and aversion
442 to disadvantageous inequity increases with the recipient's perceived competence (but
443 does not depend on their perceived warmth)⁷. In that study, the participant decided how
444 many tokens to share with the recipient in a continuous manner, and thus it was up to

445 them whether and how often they created advantageous or disadvantageous inequity.
446 We adopted a different task design in the current study, which used two-alternative
447 forced choices regarding advantageous and disadvantageous inequity in separate trials,
448 which allowed us to test the dissociable effects of perceived warmth and competence
449 on inequity preference even more directly.

450
451 We counted how often the participants chose the unequal allocation over the equal
452 allocation against each recipient in advantageous and disadvantageous inequity trials
453 and tested their correlation with the perceived warmth and competence of the recipients
454 for those choices (Fig. 1c). The statistical significance of the correlation was assessed
455 via permutation (9,999 iterations). The same permutation test was also used to assess
456 whether the effects of warmth and competence on choice frequencies were different
457 from each other (i.e., statistical significance on the difference in correlations). While Fig.
458 1c shows choice frequencies marginalized over payoff structures in each trial type, the
459 relationship with trait perceptions was robustly observed even when measured for each
460 payoff structure separately (Fig. S2a, b).

461
462 **MRI data acquisition** MR images were acquired by a 3T Siemens Magnetom Trio
463 scanner and a 12-channel head coil. A 3D high-resolution structural image was
464 acquired using a T1-weighted magnetization-prepared rapid-acquisition gradient-echo
465 (MPRAGE) pulse sequence (voxel size = 1 × 1 × 1 mm, matrix size = 190 × 239, 200
466 axial slices, TR = 2300 msec, TE = 2.98 msec). While participants completed the task,
467 functional images were acquired using a T2*-weighted gradient echo-planar imaging
468 (EPI) pulse sequence (voxel size = 3 × 3 × 3 mm, interslice gap = 0.15 mm, matrix size
469 = 64 × 64, 32 oblique axial slices, TR = 2000 msec, TE = 30 msec). Slices were angled
470 +30 degrees with respect to the anterior commissure-posterior commissure line to
471 reduce signal dropout in the orbitofrontal cortex⁹².

472
473 **MRI data analysis: trait perception.** We conducted a whole-brain searchlight
474 Representational Similarity Analysis (RSA) to look for neural representations of the
475 recipient's perceived traits⁴¹. More specifically, we looked for brain regions in which
476 voxel-wise local response patterns evoked by two recipients are similar (dissimilar)
477 when their perceived traits are also similar (dissimilar) to each other. Our RSA
478 formulated this relationship as the correlation between two representational dissimilarity
479 matrices (RDMs), one that captures dissimilarity in trait perception (trait RDM) and one
480 that captures dissimilarity in response patterns (neural RDM), in all possible pairs of
481 recipients (20 recipients, 190 pairwise similarity measures).

482
483 For the trait RDM, pairwise dissimilarity in perceived traits was quantified as Euclidean
484 distance in a two-dimensional space of perceived warmth and competence (Fig. 1a).
485 Empirical measures of warmth and competence perceptions were originally obtained as
486 numeric scores between 0 and 100⁷. We z-scored each dimension across the 20
487 recipients to construct the Euclidean space.

488

489 The neural RDM was computed at every voxel within grey matter in native space.
490 Pairwise dissimilarity in voxel-wise response patterns was quantified as the cross-
491 validated Mahalanobis (Crossnobis) distance in a gray-matter spherical searchlight
492 (10mm radius). Crossnobis distance is an unbiased measure of the extent to which
493 response patterns evoked by two recipients are *consistently distinguishable across*
494 *scanning runs*⁴². We chose this distance measure over alternative measures because
495 we were primarily interested in how recipients are *distinguished* in their neural
496 representation, rather than how they are *similarly represented*. In our experiment, since
497 each recipient was presented four times in each of the two scanning runs, we were able
498 to cross-validate distance estimates across runs to mitigate spurious distance caused
499 by noise (overfitting).

500
501 The pairwise Crossnobis distance was estimated following the formulae provided
502 previously⁴². We first estimated voxel-wise response patterns evoked by each recipient
503 in each scanning run using a GLM implemented in SPM12. To retain fine-grained
504 signals as much as possible, minimal preprocessing (only motion correction) was
505 applied to EPIs prior to the GLM. The GLM included the regressors of interest, modeling
506 the presentation of each recipient using a box-car function that starts with the onset of
507 the recipient presentation and ends with the onset of payoffs presentation, along with
508 nuisance regressors modeling button presses. These regressors were convolved with
509 the canonical double-gamma hemodynamic response function (HRF) and its temporal
510 derivative. The GLM also included confound regressors for head motion (3 translations
511 and 3 rotations, estimated in the motion correction procedure), 128-sec high-pass
512 filtering, and AR(1) model of serial autocorrelation. The GLM coefficients of each
513 recipient within the searchlight were then cross-validated across the two runs to obtain
514 the Crossnobis distance. For Mahalanobis whitening, we estimated the covariance
515 matrix in the searchlight using the GLM residuals and shrank it for invertibility⁹³.

516
517 We computed Fisher-transformed Spearman correlation between the trait and neural
518 RDMs at each gray-matter voxel. We discovered that the trait RDM inadvertently
519 contained information about visual features of the recipient presentation on the screen,
520 and specifically its character count. This visual confound was controlled by partialling
521 out another RDM that captured the character count. The resultant correlation map was
522 normalized to the standard MNI space based on the MPRAGE structural image of each
523 participant and spatially smoothed (Gaussian kernel FWHM = 8 mm) using SPM12. For
524 the population-level analysis, a cluster-level permutation test was conducted using FSL
525 *randomise* (threshold-free cluster enhancement [TFCE], whole-brain FWE corrected p
526 $< .05$, 4,999 iterations).

527
528 **MRI data analysis: correlation with individual choices.** To look for evidence that any
529 of the regions that represented the perceived traits (Fig. 2b) contributed to the
530 subsequent monetary allocation decisions, we ran another RSA which tested the
531 correlation between neural RDMs and choice RDMs. We predicted that, if a region
532 contributed to the decisions, local response patterns evoked by two recipients in one

533 participant's brain would be similar (dissimilar) to each other when the participant
534 treated them in a similar (dissimilar) manner in their allocation choices.

535
536 The individual choice RDM was built on the frequency at which each participant chose
537 the advantageous or disadvantageous unequal allocation for each recipient. Pairwise
538 Euclidean distance was measured in the two-dimensional space of the observed choice
539 frequencies, one dimension for advantageous inequity trials and the other dimension for
540 disadvantageous inequity trials. Since each recipient was presented in three
541 advantageous inequity trials and three disadvantageous inequity trials, the choice
542 frequency on each dimension was either 0, 1/3, 2/3, or 1.

543
544 These individual-level choice RDM were then correlated with neural RDMs in the
545 regions identified by our first RSA as containing representations of others' traits. Binary
546 masks were functionally defined in standard MNI space based on the aforementioned
547 population-level statistics (TFCE, whole-brain FWE corrected $p < .05$) and converted to
548 the native space of each participant's brain using SPM12. The z-transformed Spearman
549 correlation between the choice and neural RDMs was averaged across all voxels in the
550 native-space masks.

551
552 In order to test whether neural response patterns predicted individual choice patterns
553 *above and beyond* the population-level effects of warmth and competence, we
554 conducted a permutation test, randomly pairing choice and neural RDMs from different
555 participants (4,999 iterations). To control for multiple comparisons across ROIs, the null-
556 hypothesis distribution was constructed by taking the highest population average of
557 correlation scores across the ROIs in each permutation iteration.

558 **References**

- 559
- 560 1. Greenwald, A. G. & Banaji, M. R. Implicit social cognition: Attitudes, self-esteem, and
561 stereotypes. *Psychol Rev* **102**, 4–27 (1995).
- 562 2. Asch, S. E. Forming impressions of personality. *J Abnorm Soc Psychology* **41**, 258–
563 290 (1946).
- 564 3. Greenwald, A. G. & Lai, C. K. Implicit Social Cognition. *Annu Rev Psychol* **71**, 1–27
565 (2019).
- 566 4. Ames, D. R. Inside the Mind Reader’s Tool Kit: Projection and Stereotyping in Mental
567 State Inference. *J Pers Soc Psychol* **87**, 340–353 (2004).
- 568 5. Bertrand, M. & Duflo, E. Field Experiments on Discrimination. in *Handbook of Field*
569 *Experiments* (eds. Banerjee, A. & Duflo, E.) (2017).
- 570 6. Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A. & Yzerbyt, V. Navigating the social
571 world: Toward an integrated framework for evaluating self, individuals, and groups.
572 *Psychol Rev* **128**, 290–314 (2021).
- 573 7. Jenkins, A. C., Karashchuk, P., Zhu, L. & Hsu, M. Predicting human behavior toward
574 members of different social groups. *P Natl Acad Sci Usa* **115**, 9696–9701 (2018).
- 575 8. Fiske, S. T., Cuddy, A. J. C. & Glick, P. Universal dimensions of social cognition:
576 warmth and competence. *Trends Cogn Sci* **11**, 77–83 (2007).
- 577 9. Andreoni, J. & Miller, J. Giving According to GARP: An Experimental Test of the
578 Consistency of Preferences for Altruism. *Econometrica* **70**, 737–753 (2002).
- 579 10. Charness, G. & Rabin, M. Understanding Social Preferences with Simple Tests. *Q J*
580 *Econ* **117**, 817–869 (2002).
- 581 11. Fehr, E. & Schmidt, K. M. A Theory of Fairness, Competition, and Cooperation. *Q J*
582 *Econ* **114**, 817–868 (1999).
- 583 12. Tamir, D. I., Thornton, M. A., Contreras, J. M. & Mitchell, J. P. Neural evidence that
584 three dimensions organize mental state representation: Rationality, social impact, and
585 valence. *Proc Natl Acad Sci USA* **113**, 194–199 (2016).
- 586 13. Quadflieg, S. *et al.* Exploring the Neural Correlates of Social Stereotyping. *J*
587 *Cognitive Neurosci* **21**, 1560–1570 (2009).
- 588 14. Mitchell, J. P., Ames, D. L., Jenkins, A. C. & Banaji, M. R. Neural correlates of
589 stereotype application. *J Cognitive Neurosci* **21**, 594–604 (2009).
- 590 15. Schurz, M., Radua, J., Aichhorn, M., Richlan, F. & Perner, J. Fractionating theory of
591 mind: A meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev* **42**,
592 9–34 (2014).
- 593 16. Saxe, R. & Kanwisher, N. People thinking about thinking people: The role of the
594 temporo-parietal junction in “theory of mind.” *Neuroimage* **19**, 1835–1842 (2003).

- 595 17. Amodio, D. M. & Frith, C. D. Meeting of minds: the medial frontal cortex and social
596 cognition. *Nat Rev Neurosci* **7**, 268–277 (2006).
- 597 18. Spreng, R. N., Mar, R. A. & Kim, A. S. N. The Common Neural Basis of
598 Autobiographical Memory, Propection, Navigation, Theory of Mind, and the Default
599 Mode: A Quantitative Meta-analysis. *J Cognitive Neurosci* **21**, 489–510 (2009).
- 600 19. Frith, C. D. & Frith, U. Interacting Minds--A Biological Basis. *Science* **286**, 1692–
601 1695 (1999).
- 602 20. Jenkins, A. C. & Mitchell, J. P. How Has Cognitive Neuroscience Contributed to
603 Social Psychological Theory? in *Social Neuroscience: Towards Understanding the*
604 *Underpinnings of the Social Mind* (eds. Todorov, A., Fiske, S. & Prentice, D.) (Oxford
605 University Press, 2011).
- 606 21. Molenberghs, P., Johnson, H., Henry, J. D. & Mattingley, J. B. Understanding the
607 minds of others: A neuroimaging meta-analysis. *Neurosci Biobehav Rev* **65**, 276–291
608 (2016).
- 609 22. Mars, R. B. *et al.* On the relationship between the “default mode network” and the
610 “social brain.” *Front Hum Neurosci* **6**, 189 (2012).
- 611 23. Contreras, J. M., Banaji, M. R. & Mitchell, J. P. Dissociable neural correlates of
612 stereotypes and other forms of semantic knowledge. *Soc Cogn Affect Neur* **7**, 764–770
613 (2012).
- 614 24. Van der Cruyssen, L., Heleven, E., Ma, N., Vandekerckhove, M. & Van Overwalle,
615 F. Distinct neural correlates of social categories and personality traits. *Neuroimage* **104**,
616 336–346 (2015).
- 617 25. Contreras, J. M., Schirmer, J., Banaji, M. R. & Mitchell, J. P. Common Brain
618 Regions with Distinct Patterns of Neural Responses during Mentalizing about Groups
619 and Individuals. *J Cognitive Neurosci* **25**, 1406–1417 (2013).
- 620 26. Delplanque, J., Heleven, E. & Van Overwalle, F. Neural representations of Groups
621 and Stereotypes using fMRI repetition suppression. *Sci Rep-uk* **9**, 3190 (2019).
- 622 27. Harris, L. T. & Fiske, S. T. Dehumanizing the Lowest of the Low: Neuroimaging
623 Responses to Extreme Out-Groups. *Psychol Sci* **17**, 847–853 (2006).
- 624 28. Rangel, A. & Hare, T. Neural computations associated with goal-directed choice.
625 *Curr Opin Neurobiol* **20**, 262–270 (2010).
- 626 29. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: A coordinate-based
627 meta-analysis of BOLD fMRI experiments examining neural correlates of subjective
628 value. *Neuroimage* **76**, 412–427 (2013).
- 629 30. Tricomi, E., Rangel, A., Camerer, C. F. & O’Doherty, J. P. Neural evidence for
630 inequality-averse social preferences. *Nature* **463**, 1089–1091 (2010).
- 631 31. Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E. & Behrens, T.
632 E. Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron* **70**, 1054
633 1069 (2011).

- 634 32. Yu, R., Calder, A. J. & Mobbs, D. Overlapping and distinct representations of
635 advantageous and disadvantageous inequality. *Hum Brain Mapp* **35**, 3290–3301
636 (2014).
- 637 33. Ruff, C. C. & Fehr, E. The neurobiology of rewards and values in social decision
638 making. *Nat Rev Neurosci* **15**, 549–62 (2014).
- 639 34. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human Orbitofrontal Cortex
640 Represents a Cognitive Map of State Space. *Neuron* **91**, 1402–1412 (2016).
- 641 35. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a
642 cognitive map of task space. *Neuron* **81**, 267–279 (2014).
- 643 36. Padoa-Schioppa, C. & Conen, K. E. Orbitofrontal Cortex: A Neural Circuit for
644 Economic Decisions. *Neuron* **96**, 736–754 (2017).
- 645 37. Wikenheiser, A. M. & Schoenbaum, G. Over the river, through the woods: cognitive
646 maps in the hippocampus and orbitofrontal cortex. *Nat Rev Neurosci* **17**, 513–523
647 (2016).
- 648 38. Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does
649 not do. *Nat Neurosci* **18**, 620–627 (2015).
- 650 39. Niv, Y. Learning task-state representations. *Nat Neurosci* **22**, 1544–1553 (2019).
- 651 40. Yu, L. Q., Kan, I. P. & Kable, J. W. Beyond a rod through the skull: A systematic
652 review of lesion studies of the human ventromedial frontal lobe. *Cognitive Neuropsych*
653 **37**, 1–45 (2019).
- 654 41. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis—
655 connecting the branches of systems neuroscience. *Frontiers Syst Neurosci* **2**, 4 (2008).
- 656 42. Walther, A. *et al.* Reliability of dissimilarity measures for multi-voxel pattern analysis.
657 *Neuroimage* **137**, 188–200 (2016).
- 658 43. Gallagher, M., McMahan, R. W. & Schoenbaum, G. Orbitofrontal Cortex and
659 Representation of Incentive Value in Associative Learning. *J Neurosci* **19**, 6610–6614
660 (1999).
- 661 44. Jones, J. L. *et al.* Orbitofrontal Cortex Supports Behavior and Learning Using
662 Inferred But Not Cached Values. *Science* **338**, 953–956 (2012).
- 663 45. Izquierdo, A., Suda, R. K. & Murray, E. A. Bilateral Orbital Prefrontal Cortex Lesions
664 in Rhesus Monkeys Disrupt Choices Guided by Both Reward Value and Reward
665 Contingency. *J Neurosci* **24**, 7540–7548 (2004).
- 666 46. West, E. A., DesJardin, J. T., Gale, K. & Malkova, L. Transient Inactivation of
667 Orbitofrontal Cortex Blocks Reinforcer Devaluation in Macaques. *J Neurosci* **31**, 15128–
668 15135 (2011).
- 669 47. Reber, J. *et al.* Selective impairment of goal-directed decision-making following
670 lesions to the human ventromedial prefrontal cortex. *Brain* **140**, 1743–1756 (2017).

- 671 48. Wang, F., Howard, J. D., Voss, J. L., Schoenbaum, G. & Kahnt, T. Targeted
672 Stimulation of an Orbitofrontal Network Disrupts Decisions Based on Inferred, Not
673 Experienced Outcomes. *J Neurosci* **40**, 8726–8733 (2020).
- 674 49. Howard, J. D. *et al.* Targeted Stimulation of Human Orbitofrontal Networks Disrupts
675 Outcome-Guided Behavior. *Curr Biol* **30**, 490-498.e4 (2020).
- 676 50. Chan, S. C. Y., Niv, Y. & Norman, K. A. A Probability Distribution over Latent
677 Causes, in the Orbitofrontal Cortex. *J Neurosci* **36**, 7817–7828 (2016).
- 678 51. Zhou, J. *et al.* Evolving schema representations in orbitofrontal ensembles during
679 learning. *Nature* **590**, 606–611 (2021).
- 680 52. Stalnaker, T. A., Raheja, N. & Schoenbaum, G. Orbitofrontal State Representations
681 Are Related to Choice Adaptations and Reward Predictions. *J Neurosci* **41**, 1941–1951
682 (2021).
- 683 53. Nassar, M. R., McGuire, J. T., Ritz, H. & Kable, J. W. Dissociable Forms of
684 Uncertainty-Driven Representational Change Across the Human Brain. *J Neurosci* **39**,
685 1688–1698 (2019).
- 686 54. Saez, I. *et al.* Encoding of Multiple Reward-Related Computations in Transient and
687 Sustained High-Frequency Activity in Human OFC. *Curr Biol* **28**, 2889-2899.e3 (2018).
- 688 55. Gilbert, S. J., Swencionis, J. K. & Amodio, D. M. Evaluative vs. trait representation
689 in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal
690 cortex. *Neuropsychologia* **50**, 3600–3611 (2012).
- 691 56. Xia, C., Stolle, D., Gidengil, E. & Fellows, L. K. Lateral Orbitofrontal Cortex Links
692 Social Impressions to Political Choices. *J Neurosci* **35**, 8507–8514 (2015).
- 693 57. Jenkins, A. C. & Mitchell, J. P. Mentalizing under uncertainty: dissociated neural
694 responses to ambiguous and unambiguous mental state inferences. *Cereb Cortex* **20**,
695 404–10 (2010).
- 696 58. Karmarkar, U. R. & Jenkins, A. C. Neural and Behavioral Insights into Online Trust
697 and Uncertainty. in *Organizational Neuroethics* (eds. Martineau, J. & Racine, E.) 191–
698 207 (2020).
- 699 59. Carmichael, S. T. & Price, J. L. Sensory and premotor connections of the orbital and
700 medial prefrontal cortex of macaque monkeys. *J Comp Neurol* **363**, 642–664 (1995).
- 701 60. Saleem, K. S., Kondo, H. & Price, J. L. Complementary circuits connecting the
702 orbital and medial prefrontal networks with the temporal, insular, and opercular cortex in
703 the macaque monkey. *J Comp Neurol* **506**, 659–693 (2008).
- 704 61. Carmichael, S. T. & Price, J. L. Connectional networks within the orbital and medial
705 prefrontal cortex of macaque monkeys. *J Comp Neurol* **371**, 179–207 (1996).
- 706 62. Zald, D. H. *et al.* Meta-Analytic Connectivity Modeling Reveals Differential
707 Functional Connectivity of the Medial and Lateral Orbitofrontal Cortex. *Cereb Cortex* **24**,
708 232–248 (2014).

- 709 63. Kahnt, T., Chang, L. J., Park, S. Q., Heinzle, J. & Haynes, J.-D. Connectivity-based
710 parcellation of the human orbitofrontal cortex. *J Neurosci* **32**, 6240–6250 (2012).
- 711 64. Ballesta, S., Shi, W., Conen, K. E. & Padoa-Schioppa, C. Values encoded in
712 orbitofrontal cortex are causally related to economic choices. *Nature* **588**, 450–453
713 (2020).
- 714 65. Birman, D. & Gardner, J. L. A flexible readout mechanism of human sensory
715 representations. *Nat Commun* **10**, 3500 (2019).
- 716 66. Amodio, D. M. Social Cognition 2.0: An Interactive Memory Systems Account.
717 *Trends Cogn Sci* **23**, 21–33 (2019).
- 718 67. Amodio, D. M. & Cikara, M. The Social Neuroscience of Prejudice. *Annu Rev*
719 *Psychol* **72**, 1–31 (2020).
- 720 68. Olson, I. R., McCoy, D., Klobusicky, E. & Ross, L. A. Social cognition and the
721 anterior temporal lobes: a review and theoretical framework. *Soc Cogn Affect Neur* **8**,
722 123–133 (2013).
- 723 69. Tavares, R. M. *et al.* A Map for Social Navigation in the Human Brain. *Neuron* **87**,
724 231–243 (2015).
- 725 70. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible
726 Behavior. *Neuron* **100**, 490–509 (2018).
- 727 71. Park, S. A., Miller, D. S., Nili, H., Ranganath, C. & Boorman, E. D. Map Making:
728 Constructing, Combining, and Inferring on Abstract Cognitive Maps. *Neuron* **107**, 1226-
729 1238.e8 (2020).
- 730 72. Kubota, J. T., Banaji, M. R. & Phelps, E. A. The neuroscience of race. *Nat Neurosci*
731 **15**, 940–948 (2012).
- 732 73. Mattan, B. D., Wei, K. Y., Cloutier, J. & Kubota, J. T. The Social Neuroscience of
733 Race- and Status-Based Prejudice. *Curr Opin Psychology* **24**, 27–34 (2018).
- 734 74. Amodio, D. M. The neuroscience of prejudice and stereotyping. *Nat Rev Neurosci*
735 **15**, 670–682 (2014).
- 736 75. Gray, K., Jenkins, A. C., Heberlein, A. S. & Wegner, D. M. Distortions of mind
737 perception in psychopathology. *Proc National Acad Sci* **108**, 477–479 (2011).
- 738 76. Van Overwalle, F. & Vandekerckhove, M. Implicit and explicit social mentalizing:
739 dual processes driven by a shared neural network. *Front Hum Neurosci* **7**, 560 (2013).
- 740 77. Schurz, M. *et al.* Toward a hierarchical model of social cognition: A neuroimaging
741 meta-analysis and integrative review of empathy and theory of mind. *Psychol Bull* **147**,
742 293–327 (2021).
- 743 78. Van Overwalle, F., Ma, N. & Baetens, K. Nice or nerdy? The neural representation
744 of social and competence traits. *Soc Neurosci* **11**, 1–12 (2015).
- 745 79. Li, M. *et al.* Warmth is more influential than competence: an fMRI repetition
746 suppression study. *Brain Imaging Behav* **15**, 266–275 (2021).

- 747 80. Van Overwalle, F. Social cognition and the brain: A meta-analysis. *Hum Brain Mapp*
748 **30**, 829–858 (2009).
- 749 81. Heleven, E. & Van Overwalle, F. The person within: memory codes for persons and
750 traits using fMRI repetition suppression. *Soc Cogn Affect Neur* **11**, 159–171 (2016).
- 751 82. Mitchell, J. P., Macrae, C. N. & Banaji, M. R. Dissociable Medial Prefrontal
752 Contributions to Judgments of Similar and Dissimilar Others. *Neuron* **50**, 655–663
753 (2006).
- 754 83. Jenkins, A. C., Macrae, C. N. & Mitchell, J. P. Repetition suppression of
755 ventromedial prefrontal activity during judgments of self and others. *Proc Natl Acad Sci*
756 *USA* **105**, 4507–4512 (2008).
- 757 84. Mitchell, J. P., Banaji, M. R. & Macrae, C. N. The Link between Social Cognition and
758 Self-referential Thought in the Medial Prefrontal Cortex. *J Cognitive Neurosci* **17**, 1306–
759 1315 (2005).
- 760 85. Boccadoro, S. *et al.* Defining the neural correlates of spontaneous theory of mind
761 (ToM): An fMRI multi-study investigation. *Neuroimage* **203**, 116193 (2019).
- 762 86. Kestemont, J., Vandekerckhove, M., Ma, N., Hoeck, N. V. & Van Overwalle, F.
763 Situation and person attributions under spontaneous and intentional instructions: an
764 fMRI study. *Soc Cogn Affect Neur* **8**, 481–493 (2013).
- 765 87. Tamir, D. I. & Mitchell, J. P. Anchoring and adjustment during social inferences. *J*
766 *Exp Psychology Gen* **142**, 151–162 (2013).
- 767 88. Hackel, L. M., Doll, B. B. & Amodio, D. M. Instrumental learning of traits versus
768 rewards: dissociable neural correlates and effects on choice. *Nat Neurosci* **18**, 1233–
769 1235 (2015).
- 770 89. Mende-Siedlecki, P., Cai, Y. & Todorov, A. The neural dynamics of updating person
771 impressions. *Soc Cogn Affect Neur* **8**, 623–631 (2013).
- 772 90. Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C. & Fehr, E. Linking brain structure
773 and activation in temporoparietal junction to explain the neurobiology of human altruism.
774 *Neuron* **75**, 73–79 (2012).
- 775 91. Bruhin, A., Fehr, E. & Schunk, D. The many faces of human sociality: Uncovering
776 the distribution and stability of social preferences. *J Eur Econ Assoc* **72**, 738 (2018).
- 777 92. Weiskopf, N., Hutton, C., Josephs, O. & Deichmann, R. Optimal EPI parameters for
778 reduction of susceptibility-induced BOLD sensitivity losses: A whole-brain analysis at
779 3 T and 1.5 T. *Neuroimage* **33**, 493–504 (2006).
- 780 93. Ledoit, O. & Wolf, M. Honey, I Shrank the Sample Covariance Matrix. *J Portfolio*
781 *Management* **30**, 110–119 (2004).

782

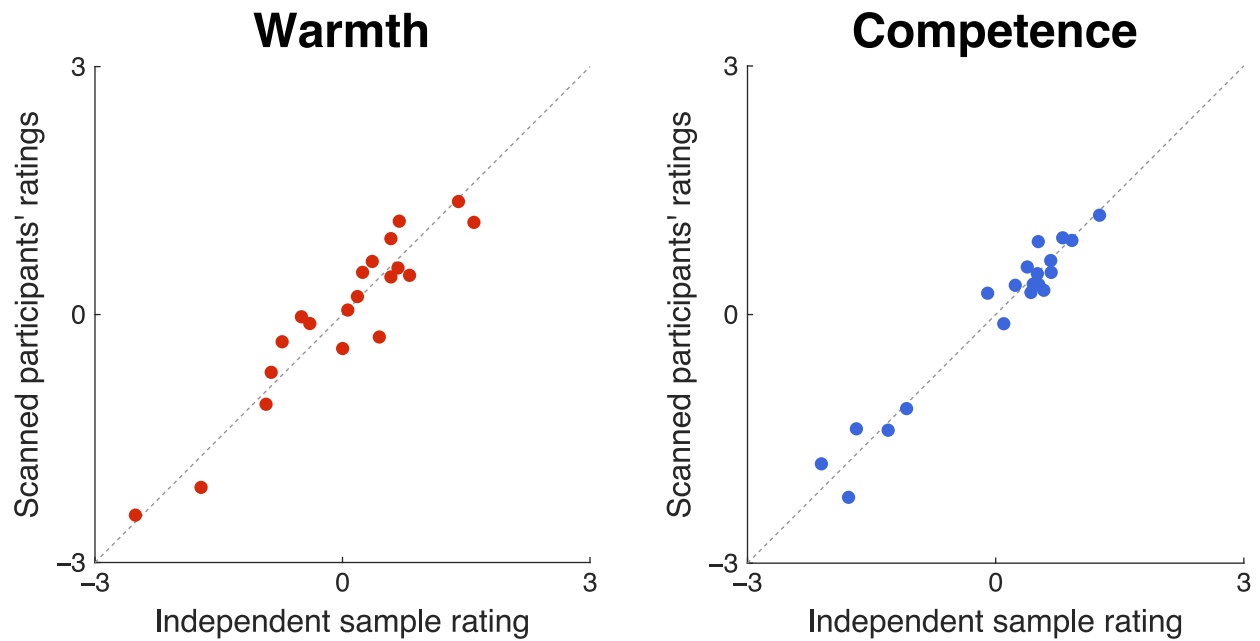
783 **Acknowledgments**

784

785 The authors thank Nakyung Lee and Pierre Karashchuk for assistance with paradigm
786 development; Duy Phan, Amanda Savarese, and Cassandra Carrin for assistance with
787 data collection; and Dilara Berkay for helpful input and assistance with the preparation
788 of fMRI data for analysis.

789 **Supplementary Figures**

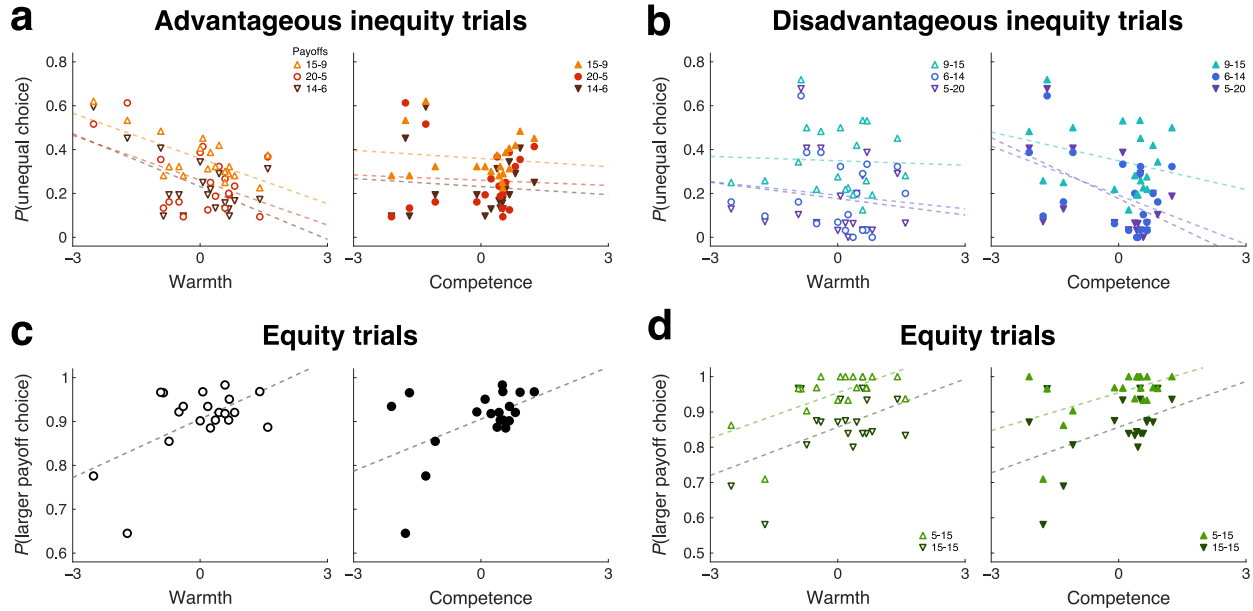
790



791

792

793 **Fig. S1.** Consistency in trait perception. In all behavioral and fMRI analyses, we used
794 ratings of warmth and competence from our previous study (Jenkins et al., 2018, Study
795 1b, $n = 252$; x axis). We also collected ratings from our participants after scanning ($n =$
796 32 ; y axis). These two sets of ratings are highly correlated (warmth: Pearson's $r = .943$,
797 competence: $r = .978$), demonstrating the robustness of trait perceptions.



798
799

800 **Fig. S2.** The effect of perceived traits on monetary allocation choices, separately for
 801 each payoff structure. **a.** In advantageous inequity trials, the unequal self-recipient
 802 allocations were either \$15-\$9, \$20-\$6, or \$14-\$6. Consistent patterns were observed
 803 across these payoff conditions; participants were less likely to choose the unequal
 804 allocation as the recipient's perceived warmth was higher (*left*, 15-9: Pearson's $r = -.68$,
 805 permutation $p = .001$, 20-5: $r = -.47$, $p = .021$, 14-6: $r = -.60$, $p = .004$) irrespective of
 806 the recipient's perceived competence (*right*, 15-9: $r = -.12$, $p = .285$, 20-5: $r = -.06$, p
 807 $= .386$, 14-6: $r = -.09$, $p = .331$), and the effect of warmth was stronger than
 808 competence (15-9: $p = .001$, 20-5: $p = .017$, 14-6: $p = .004$). **b.** In disadvantageous
 809 inequity trials, the unequal self-recipient allocations were either \$9-\$15, \$6-\$14, or \$5-
 810 \$20. Consistent patterns were observed across these payoff conditions, except that the
 811 competence effect did not reach statistical significance in 9-15; participants were less
 812 likely to choose the unequal allocation as the recipient's perceived competence was
 813 higher (*right*, 9-15: $r = -.28$, $p = .125$, 6-14: $r = -.44$, $p = .036$, 5-20: $r = -.52$, $p = .018$)
 814 irrespective of the recipient's perceived warmth (*left*, 9-15: $r = -.04$, $p = .417$, 6-14: $r =$
 815 $-.12$, $p = .287$, 5-20: $r = -.14$, $p = .265$), and the effect of competence was stronger than
 816 warmth (9-15: $p = .120$, 6-14: $p = .054$, 5-20: $p = .024$). **c.** In some trials, the participant
 817 was presented with two equal allocations (one option was \$10-\$10, and the other option
 818 was either \$5-\$5 or \$15-\$15). These conditions were only included to encourage the
 819 participant to pay attention to both sets of payoffs and were not discussed in the main
 820 text. In these trials, participants chose the option with higher payoffs more often when
 821 the recipient's warmth was higher ($r = .57$, $p = .009$), and also when their competence
 822 was higher ($r = .51$, $p = .022$). The effects of warmth and competence did not differ
 823 significantly ($p = .362$). These results demonstrate that participants incorporated the
 824 recipient's warmth and competence into their choices in a highly context-dependent
 825 manner. **d.** Consistent behavioral patterns were observed across both payoff conditions
 826 in the equity trials; the larger payoff frequency increased with warmth (*right*, 5-5: $r = .63$,

827 $p = .006$, 15-15: $r = .49$, $p = .022$) and competence (*left*, 5-5: $r = .52$, $p = .020$, 15-15: r
828 $= .46$, $p = .033$), and their effects were comparable (5-5: $p = .287$, 15-15: $p = .440$).