COGNITIVE NEUROSCIENCE

Domain-specific representation of social inference by neurons in the human amygdala and hippocampus

Runnan Cao[1]*, Julien Dubois[2], Adam N. Mamelak[2], Ralph Adolphs[3,4]†, Shuo Wang[1]†, Ueli Rutishauser[2,4]*†

Inferring the intentions and emotions of others from behavior is crucial for social cognition. While neuroimaging studies have identified brain regions involved in social inference, it remains unknown whether performing social inference is an abstract computation that generalizes across different stimulus categories or is specific to certain stimulus domain. We recorded single-neuron activity from the medial temporal lobe (MTL) and the medial frontal cortex (MFC) in neurosurgical patients performing different types of inferences from images of faces, hands, and natural scenes. Our findings indicate distinct neuron populations in both regions encoding inference type for social (faces, hands) and nonsocial (scenes) stimuli, while stimulus category was itself represented in a task-general manner. Uniquely in the MTL, social inference type was represented by separate subsets of neurons for faces and hands, suggesting a domain-specific representation. These results reveal evidence for specialized social inference processes in the MTL, in which inference representations were entangled with stimulus type as expected from a domain-specific process.

INTRODUCTION

Inferring the latent mental states and intentions of other people from observing their behavior is a critical human ability that is at the core of what is often referred to as either “theory of mind” (ToM), “mentalizing” (*1*, *2*), or “social inference” (*3*, *4*). Emerging during early childhood, typically around the age of four (*5*), this important skill relies on representing the beliefs, desires, intentions, and feelings of others to make sophisticated social interactions possible (*6*). Atypical social inference is thought to contribute to the difficulties experienced in mental and neurological disorders, including in autism (*7*, *8*), schizophrenia (*9*), and Parkinson’s disease (*10*, *11*) [for review, see (*12*)]. Social inference has been an active topic of study since the 1970s, ranging from developmental psychology (*13*) to social neuroscience (*14*, *15*) to philosophy of mind (*16*). A wide range of tasks has been developed to study it, including the false-belief task (*17*), pragmatic language comprehension (*18*), and belief-desire reasoning (*19*). A key question in social inference work is how the brain represents its own and other’s mental states at the neuronal level (*6*).

Anatomically, a large number of brain regions have been associated with social inference, ranging from the cerebellum to the superior temporal sulcus to the frontal cortex (*12*, *20*, *21*), a list that varies greatly depending on the exact task used [see (*22*, *23*) for reviews]. Three common sets of regions that stand out in the literature are the temporo-parietal junction (TPJ) (*24*, *25*), medial frontal cortex (MFC) (*21*), and the medial temporal lobe (MTL), which together form important components of the “social brain” (*26*, *27*). The MFC contains several areas of interest to social inference, including the supplementary motor area (SMA), the pre-SMA, the anterior and middle cingulate cortex (ACC/MCC) (*28*), and the medial prefrontal cortex (mPFC) (*29*, *30*). These subregions have been broadly associated, to varying extents, with the inference of different mental states (*31*), such as false beliefs (*32*), deception (*20*), intentions (*33*), empathy (*34*), desires (*24*), and preferences (*35*), indicating a prominent role for sectors of the MFC in specific kinds of social inferences (*36*). The role of frontal regions in social inference is dissociable from their role in “executive functions” more broadly (*37*, *38*), suggesting that social inference processes are distinct specializations rather than reutilization of more general executive processes.

By contrast, a broader role in making social inferences is suggested for the MTL, notably including the amygdala (AMY) and hippocampus (HIPP) (*26*, *27*, *39*). The MTL supports processes such as recognition memory (*40*), social evaluations (*27*, *41*), categorization (*42*), facial emotion recognition (*43*, *44*), relational processing (*45*), and latent state inference (*46*) that are needed for social inference but are not specialized for doing so. The MTL is closely connected both structurally and functionally with the MFC (*47*, *48*), suggesting that these two regions are two nodes in the social inference network.

Several important questions regarding the neural basis of social inference remain open. First, while a wide set of brain regions have been implicated in social inference, it remains unknown what specifically the neurons in these regions contribute functionally, a question that neuroimaging studies alone cannot resolve. Second, although many subregions of the MFC have been linked to social inference, only the mPFC reached a high degree (90%) of reliability of activation across different inference studies (*36*). It remains unclear how other frontal brain regions, including pre-SMA and dorsal anterior cingulate cortex (dACC), contribute to social inference. Third, while the MTL has been hypothesized to play an important role in social inference (*26*), it is not typically identified as part of the social inference network in imaging studies (*36*, *49*). However, lesion studies (e.g., impaired social inference function following amygdala damage) (*49*–*51*) as well as prominent representations of faces and judgments about faces at the single neuron level (*52*, *53*) suggest that the amygdala plays an important role in social inference. This discrepancy may be partly attributed to limitations such

[1]Department of Radiology, Washington University in St. Louis, St. Louis, MO 63110, USA. [2]Departments of Neurosurgery, Neurology, and Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA. [3]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, U.S.A. [4]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, U.S.A.
*Corresponding author. Email: r.cao@wustl.edu (R.C.); ueli.rutishauser@cshs.org(U.R.)
†These authors contributed equally to this work.

as poor signal-to-noise ratio (SNR) in subcortical areas with blood oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI), thus leaving the involvement of the MTL in social inference unclear.

Last, it remains unclear whether the implementation of social inference recruits specific brain regions dedicated to representing mental states (*54*, *55*). This question is part of a long standing debate on the domain specificity of social processing in the brain, with seminal debates historically focused on whether or not there are regions specialized for processing faces (*56*). With respect to the processes engaged in social inference, this important question remains unresolved. On one hand, a recent nonhuman primate fMRI study found that part of the MFC was exclusively activated during social interactions (*57*), and lesion studies in the macaque have found evidence for specifically social valuation in anterior cingulate cortex (*58*). On the other hand, human studies suggested that the MFC plays a more general role in subserving executive function–related neural functions rather than social inference specifically (*54*, *59*, *60*).

Here, we used single-neuron recordings to study social inference using a well-validated and well-established ToM task. Our focus is on the specific cognitive function of inferring the mental states of others from observing human behavior. We previously developed and validated with fMRI and behavior a task that contrasts physical judgments about social images ("how" an action is being performed) with social inferences about the mental states responsible ("why" the person is performing that action) (*61*). We note that this task has been adopted in the National Institute of Mental Health (NIMH) Research Domain Criteria framework for the subconstruct of "action perception" within the construct "perception and understanding of others" in the social processes domain of the framework (www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/action-perception) (*62*, *63*). Given the importance of this task, it is important to anchor its neural correlates in intracranial electrophysiology. Here, we used this same task with intracranial recordings in humans and asked whether social inference-related processes are represented in the responses of single neurons within parts of the MFC (the dACC and pre-SMA) and the MTL (amygdala and hippocampus). The core analysis approach we took is to contrast why with how questions for the very same images, thereby differentiating the neural representations of social inference from those of perceptual judgments (*61*).

RESULTS

Task and behavior

We used the validated "why/how" social inference task (*61*, *64*, *65*) to probe the neural mechanisms of social inference in the human brain. Patients were presented with naturalistic color images and asked to answer questions about the stimuli that required performing social inference. We varied stimulus domains (hand actions versus facial expressions versus nonsocial events) and the type of inference required to answer a question (perceptual judgment of the action probed by how questions (e.g., "is the person smiling?") versus social inference of the cause probed by why questions (e.g., "is the person admiring someone?") in a blocked design (Fig. 1A and see Materials and Methods for details). In each trial, patients were first shown the question to be answered, then saw a single image, and then made a "yes" or "no" decision. Our analysis generally used a two (inference type: why versus how) by two (category: faces versus hands) by two (choice: yes versus no) factorial design. In 8 of the 14 patients, we also included a third category of images: scenes of natural events that contained neither a face nor hands. We refer to face and hand images as social stimuli and to scene images as nonsocial stimuli. In the latter subset of sessions, the paradigm was a 2 × 3 × 2 design.

Patients' responses were compared with normative data acquired independently in (*61*). We used the data from this group of subjects as the "ground truth" to calculate the accuracy of the judgments made by the subjects in the present study. The patients performed well on both why (Fig. 1, C and D; accuracy: 87.08 ± 10.5% [mean ± SD]; response time: 0.77 ± 0.12 s [mean ± SD]) and how questions (accuracy: 91.87 ± 7.83% [mean ± SD]; response time: 0.68 ± 0.14 s [mean ± SD]). The accuracy for why questions was lower (two-tailed paired t test: $t_{18} = 4.06$, $P = 0.0007$), and the response time for why questions was longer (two-tailed paired t test: $t_{18} = 5.05$, $P = 8.03 \times 10^{-5}$) compared to how questions. This is similar to the normative data (*61*) and is expected given the additional inferential processing required for why blocks.

Neuronal correlates of social inference in the MTL and MFC

We isolated in total 726 single neurons from the amygdala, hippocampus, dACC, and pre-SMA across 19 sessions in 14 neurosurgical patients (see Fig. 1B for example locations of the electrodes and table S1 for a complete list; n = 236, 158, 141, and 191 neurons from the AMY, HIPP, dACC, and pre-SMA, respectively). For brevity, we refer to AMY and HIPP together as the MTL (n = 394 cells) and the dACC and pre-SMA together as the MFC (n = 332 cells). Only neurons with an average firing rate greater than 0.2 Hz (n = 683) were included in subsequent analyses.

Answering why and how questions required different types of inference for the very same images, thereby allowing us to isolate signatures of inference controlling for sensory input (which is the same). We first examined neural activity in a single 1-s long time window following stimulus onset (200 to 1200 ms relative to stimulus onset). A total of 17.9% of neurons in the MFC responded differentially as a function of whether subjects were performing the why or how task ("inference-type neurons"; see Fig. 2E for a summary of selected neurons in each subregion; Fig. 2, B and D shows examples; 56 of 313, 16 in dACC and 40 in pre-SMA; binomial test, $P < 10^{-20}$; three-way analysis of variance (ANOVA); see Materials and Methods). Similarly, 13.2% of MTL neurons differentiated between the two tasks following stimulus onset (Fig. 2, A and C show examples; 49 of 370, 13.2%, 28 in AMY and 21 in HIPP; binomial test,). $P = 2.58 \times 10^{-20}$ The proportion of neurons doing so was not significantly different between the MTL and MFC (17.9% versus 13.2%; χ^2 test of proportion: $P = 0.09$). Among all inference-type neurons, 60 of 105 (57.14%; see Fig. 2, A and B for examples and Fig. 2G for group results, and see fig. S1G for the proportions in each area, respectively) showed higher activity in the why task (why-preferring), and 45 (42.86%; see Fig. 2, C and D for examples and Fig. 2H for group results) had a greater response in the how task (how-preferring). The proportion of why- and how-preferring inference-type neurons was similar across brain areas and hemispheres (see fig. S1G and legend for details). As a control, we also repeated the above analysis by selecting neurons with linear regression using response time as a nuisance regressor, with qualitatively similar results (see the Supplementary Materials).

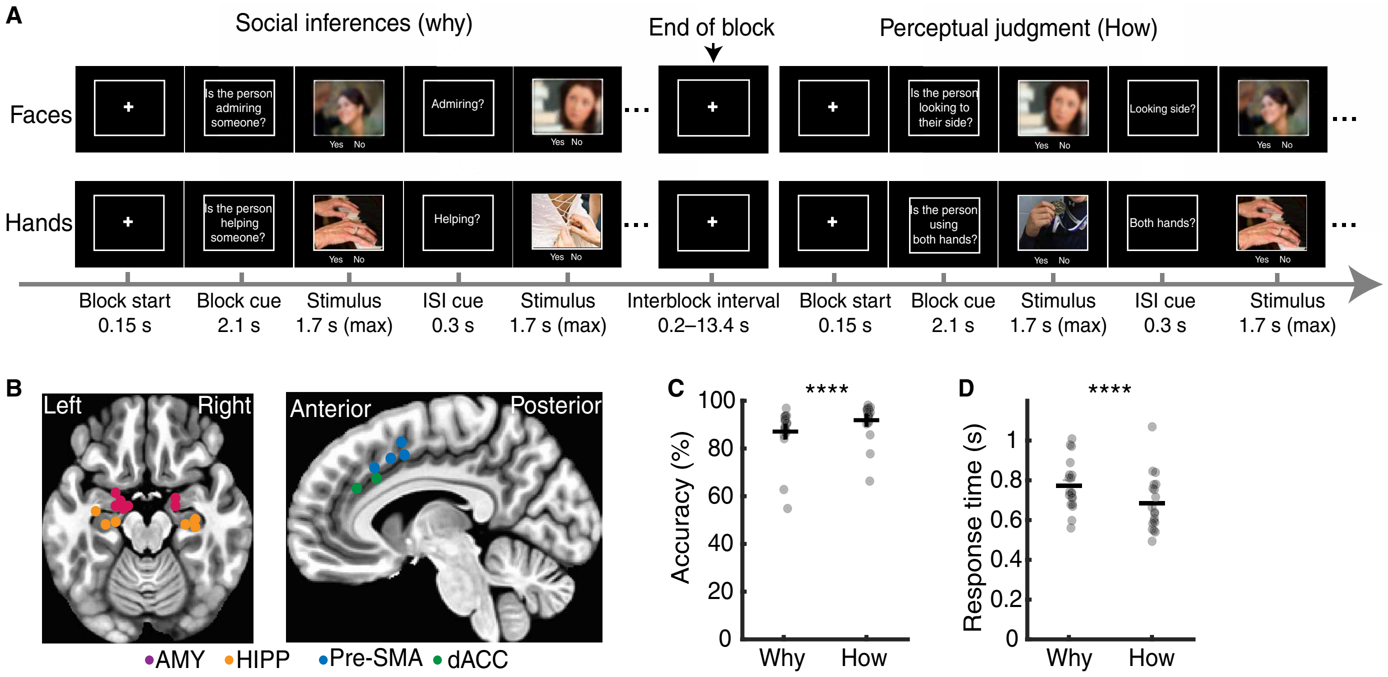

Fig. 1. Task, electrode locations, and behavior. (**A**) Task paradigm. Each session consisted of 16 blocks of 8 trials, with 128 trials in total. In each block, a set of face images with emotional expressions or hand images depicting intentional actions were paired with questions about motive (why) and implementation (how). The blocks alternated between why and how questions. Images of the same category were shown in neighboring blocks. Each block began with a short fixation and full question presentation. A brief interstimulus interval (ISI) cue was presented as a reminder of the question between image presentations. Independently acquired normative data are used to ensure that the selected images featuring unambiguous (i.e., consensus) response. Each block contained five images eliciting a yes response and three images eliciting a no response. The participants had up to 1.7 s to respond. The task advanced either 0.2 s after a response or when the response time limit was reached. The block onsets were predesigned and fixed, although the block durations were contingent on response times. As a result, session durations of were approximately equal across participants. (**B**) Example electrode locations are shown on an MNI152 template brain. Each dot indicates the location of a microwire bundle in one subject. (**C**) Behavior performance. Accuracy was calculated by comparing the participants' response to the normative response. Each dot represents a session. Only trials where participants responded were included in the analysis. (**D**) Reaction time in why versus how trials. ****$P < 0.0001$.

In addition to the above effects following stimulus onset, inference-type neurons also differentiated between task types during the interstimulus interval period that preceded stimulus onset. This was possible because the task is blocked (Fig. 1A; Fig. 2F shows an example neuron; and Fig. 2, G and H shows the average firing rate during the baseline period throughout the entire block for all inference-type neurons; and fig. S1, A to F shows the temporal dynamics aligned to trial onset).

We next investigated how the neural population as a whole represented different types of inference. We performed single-trial population decoding on the firing rates following stimulus onset (200 to 1200 ms) of all recorded neurons pooled across patients to distinguish between inference types (why versus how). Decoding accuracy was significantly above chance in both the MTL (Fig. 2I; accuracy = 73.68 ± 4.69% [mean ± SD], $P < 0.001$, compared against the empirical null distribution) and the MFC (accuracy = 86.39 ± 3.67% [mean ± SD], $P < 0.001$, compared against the empirical null distribution). Decoding accuracy in the MFC was significantly higher than that in the MTL (difference in accuracy: 12.72%, $P = 0.04$, compared against the difference of empirical null distribution), suggesting that the MFC had a stronger association with social inference at the population level (see Fig. 2K for the decoding performance in each subregion of the MTL and MFC). Similar results were derived when we matched the number of neurons between the MTL and MFC. Inference type was decodable through the whole trial including the pre-cue period (Fig. 2J and see fig. S1L for decoding performance in the MTL and MFC separately). Together, these data show that the type of inference is encoded in both MFC and MTL.

Generalizability of inference encoding in the human MTL and MFC

We next examined how the encoding of inference type (how versus why) was modulated by other task variables. We first turned to the visual category (i.e., face and hand), which is prominently encoded across the ventral visual pathway (*66*–*68*) and the MTL (*42*, *69*). In the human MFC, on the other hand, representations of visual category are task dependent in a manner that is little understood, with encoding in some and weak encoding in other tasks (*70*).

We examined whether the selectivity of inference-type neurons differed between images showing faces and hands. To do so, we selected inference-type neurons separately in face and hand trials (using a paired *t* test). A significant number of inference-type cells (Fig. 3C and see fig. S2C for the proportions in each subregion) was identified for both face (45 of 370, 12.16% in the MTL; 45 of 313, 14.38% in the MFC) and hand stimuli ($n = 31$, 8.38% in the MTL and $n = 48$, 15.34% in the MFC). Therefore, both the MTL and MFC represented inference types during both face and hand stimuli.

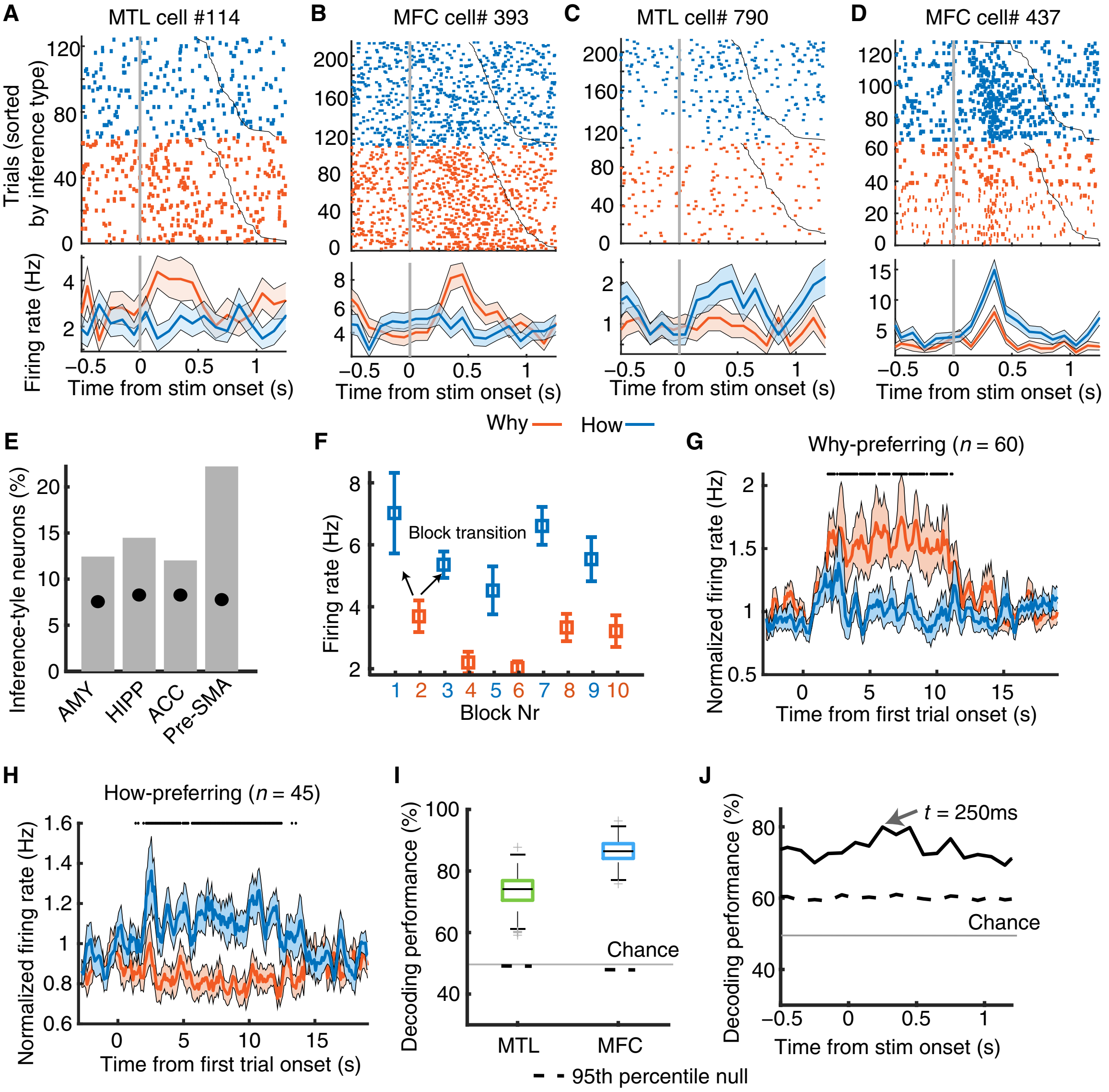

Fig. 2. Representation of inference type. (**A** to **D**) Example cells that discriminate between social (why) and perceptual (how) inference. [(A) and (C)] MTL. [(B) and (D)] MFC. (**E**) Percentage of inference-type neurons in each brain area. Gray bars represent the percentage, and black circles indicate the chance level estimated by a permutation test. (**F**) Average firing rate during the baseline period (−0.4 to 0 s relative to the stimulus onset) for each block for the cell shown in (D). Neighboring blocks for the same condition were collapsed, resulting in 10 inference-alternating blocks. (**G** and **H**) Average normalized response in why versus how blocks for all inference selective neurons. The responses were aligned to the onset of the first trial in each block. Shaded area denotes ±SEM across neurons. A dot indicates a significant difference between the conditions in that bin ($P < 0.05$, two-tailed t test, corrected by false discovery rate (90) for $Q < 0.05$, bin size = 500 ms, sliding window = 100 ms). (G) Inference-type neurons ($n = 60$) that responded more strongly to social inference (why). (H) Inference-type neurons ($n = 45$) that had a stronger responses to perceptual judgment (how). (**I** and **J**) Population decoding of inference type. The central mark on each box indicates the median, and the top and bottom edges of the box represent the 75th and 25th percentiles, respectively. "+" symbol indicates outliers. (I) Decoding with mean firing rate on all MTL neurons ($n = 370$) versus MFC neurons ($n = 313$). (J) Decoding with a sliding time window on the whole population ($n = 683$; see Materials and Methods). AMY, amygdala; HIPP, hippocampus; ACC, anterior cingulate cortex; SMA, supplementary motor area.

Strikingly, the inference-type neurons selected from trials in which faces were shown were largely distinct from those selected in trials in which hands were shown in the MTL (Fig. 3D; Chi-squared test of independence, $P = 0.82$) and the MFC, although only marginally so, indicating more overlap in the MFC compared to MTL (Fig. 3E; Chi-squared test, $P = 0.08$). This result also held in all subregions ($Ps > 0.05$) of the MTL and MFC (fig. S2C). Consistent results were revealed at the single-trial level (fig. S2, A and B). Together, this single-neuron analysis indicates that inference-type signals may generalize across face and hand stimuli in the MFC, particularly in the pre-SMA. We next tested this prediction at the population level.

We next examined how inference type was represented at the population level of all recorded neurons, with the ultimate goal of examining whether representations generalize across hands and face. We trained decoders to distinguish between how versus why questions on one visual category (i.e., face) and then tested them on

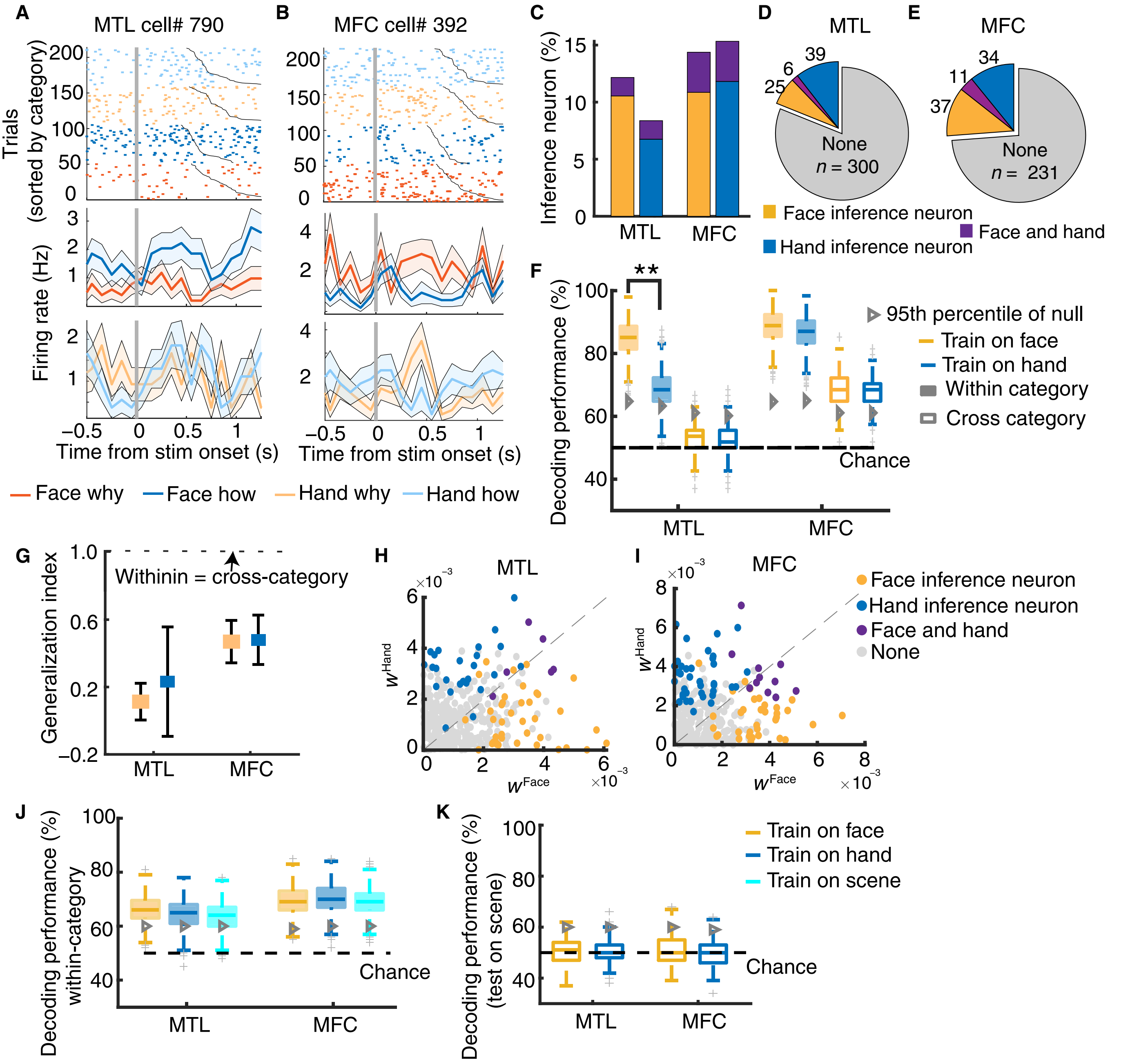

Fig. 3. Domain-specific inference type encoding. (**A** and **B**) Example inference-type cells. Inference type contrast was shown separately for face and hand stimuli, with different colors (dark colors for face and light colors for hand). (A) An example cell in the MTL that did not generalize across face and hand stimuli. (B) An example cell in the MFC that showed generalized effect across face and hand. (**C**) Proportions of inference-type neuron in the MTL and MFC. Yellow: selected with face stimuli; blue: selected with hand stimuli; purple: overlapping for face and hand. (**D** and **E**) Distribution of inference-type cells selected using face and hand stimuli respectively in the MTL (D) and MFC (E). (**F**) Population decoding of inference type using within-category decoder (i.e., train and test using either face or hand stimuli only) and cross-category decoder (i.e., train with one category and test on the other). (**G**) Generalization index of inference decoding (see Materials and Methods for computation). The representation of inference generalized across face and hand in the MFC but not in the MTL. (**H** and **I**) Scatter plot of the importance index (see Materials and Methods for details) assigned by an inference-type decoder to each cell built with face stimuli (*x* axis) or hand stimuli (*y* axis). (H) MTL cells. (I) MFC cells. (**J**) Inference-type decoders trained and tested within each category. Only cells (*n* = 281 in total) that collected in sessions where scene images presented in addition to face and hand images were included in the analysis. Legend conventions as in (G). (**K**) Decoding performance of inference type from cross social-domain decoders: train on face and test on scene (yellow) or train on hand and test on scene (blue). Legend conventions as in (G).

the same visual category (within-category decoding, i.e., face) or the other (cross-category decoding, i.e., hand). Confirming our earlier finding (for which we pooled across hands and faces), inference type was decodable using within-category decoders for both face and hand stimuli in the MTL (Fig. 3F; face: 84.40 ± 5.10% [mean ± SD]; $P < 0.001$; hand: 68.48 ± 5.91% [mean ± SD]; $P = 0.001$) and the MFC (face: 88.92 ± 4.67% [mean ± SD]; $P < 0.001$; hand: 86.48 ± 4.94% [mean ± SD]; $P < 0.001$). Decoding accuracy for face stimuli was higher than that for hand stimuli in the MTL (difference in accuracy with face-hand: 15.92%, $P = 0.01$, compared against the difference of empirical null distribution). This was the case in both the AMY (difference in accuracy with face-hand: 7.23%, $P = 0.01$) and HIPP (difference in accuracy with face-hand: 14.70%, $P = 0.01$; fig. S2D).

By contrast, decoding accuracy was not significantly different in the MFC (difference in accuracy with face-hand: 2.44%, $P = 0.08$). However, there were notable differences when looking at dACC and pre-SMA separately: Decoding accuracy was higher for face stimuli (difference in decoding accuracy face-hand: 9.91%, $P = 0.01$) in the dACC, whereas the pre-SMA had a higher decoding accuracy for hand stimuli (difference in accuracy hand-face: 3.83%, $P = 0.005$). We next turned to examine cross-condition generalization performance (train on face, test on hand, and vice versa). This revealed that, in the MFC, decoding generalized (train with faces and test with hands: 68.05 ± 4.43% [mean ± SD], $P < 0.001$; train with hands and test with faces: 67.22 ± 4.73% [mean ± SD], $P < 0.001$). In contrast, in the MTL, cross-condition generalization was not greater than expected by chance (train with faces and test with hands: 53.25 ± 4.37% [mean ± SD], $P = 0.17$; train with hands and test with faces: 52.78 ± 4.75% [mean ± SD], $P = 0.23$). This was also the case separately in both AMY and HIPP. Quantifying this observation with the generalization index (see Materials and Methods for the definition) confirmed this observation (Fig. 3G and fig. S2E). Consistently, face-selected inference-type neurons and hand-selected inference-type neurons tended to contribute exclusively to the decoding of inference type for one category in the MTL (Fig. 3H; importance index defined using weight in the decoder for each neuron) but exhibited mixed effects in the MFC (Fig. 3I). These results indicate that, in the MTL, social inference processes are coupled to specific classes of stimuli and do not generalize across stimulus categories (especially so for faces). In contrast, in the MFC, inference processes were domain-general across the two types of stimulus categories (faces and hands).

Generalizability of inference representation between social versus nonsocial world

While both MTL and MFC are implicated in social processing (*57*, *71*), they are also involved in the processing of general nonsocial objects (e.g., selectivity to different object categories) (*69*). This thus raises the question of whether making inferences in the social and nonsocial world share a common neural mechanism in these brain areas—a question related to the long-standing question about whether social processing is specialized in some way. To address this question, we also included images of scenes showing nonhuman natural events in the task in a subset of patients ($n = 281$ neurons from nine sessions; see Materials and Methods for details). As before, we asked our patients, for the same image, to either judge its perceptual properties (e.g., "is the photo showing rain?") or make inferences about the hidden states that caused what the image shows (e.g., "is it a result of thunderstorm?"). At the single-neuron level, a significant number of neurons (fig. S3, C and D) discriminated why versus how following the onset of scene images in both MTL (18 of 174, 10.34%, binomial $P = 0.0012$; 8.91% in AMY and 12.33% in HIPP; see fig. S3A for an example) and MFC (18 of 107, 16.82%, $P = 1.47 \times 10^{-6}$; 13.95% in dACC and 18.75% in pre-SMA; see fig. S3B for an example). Analysis of the single-trial response selectivity index (RSI) confirmed that these neurons discriminated why versus how questions for natural scene stimuli [fig. S3E; Kolmogorov-Smirnov (KS) test: MTL, KS = 0.21, $P = 0.59 \times 10^{-17}$; MFC, KS = 0.22, $P = 0.02 \times 10^{-17}$]. At the population level, inference type was decodable for scene images in both MTL (Fig. 3J; 63.50 ± 4.76% [mean ± SD]; $P = 0.004$) and MFC (69.03 ± 4.87% [mean ± SD]; $P = 0.004$). These results suggested that the MTL and MFC represent inference type also for nonsocial images.

We next repeated the cross-condition generalization analysis for the scene images. First, mirroring our earlier finding, inference-type neurons selected using social stimuli were largely separate from those selected using scene stimuli in both the MTL (3 of the 21 face-selected and 3 of the 14 hand-selected inference-type neurons were also selective for scene inference) and MFC (3 of the 19 face-selected and 3 of the 16 hand-selected inference-type neurons were also selective for scene inference). Second, single-trial RSI analysis confirmed this result by showing that the inference-type neurons in the MTL selected with social images could not discriminate why versus how conditions of scene stimuli (fig. S3, E and F). In line with individual neuron level results, decoding did not generalize across categories (Fig. 3K; face versus scene and hand versus scene) in neither the MTL (train with faces and test with scene: 49.87 ± 4.84% [mean ± SD], $P = 0.42$; train with hands and test with scenes: 49.97 ± 4.54% [mean ± SD], $P = 0.50$) nor MFC (train with faces and test with scene: 50.27 ± 5.93% [mean ± SD], $P = 0.44$; train with hands and test with scenes: 50.07 ± 5.43% [mean ± SD], $P = 0.54$).

Together, our results suggest that the neural substrates in the MTL and MFC for making inferences in the social versus nonsocial domain are domain specific. In contrast, in the MFC, inference was domain general between different subtypes of social domains (hands and faces).

Representation of visual categories in the MTL and MFC

An open question is whether social inference processes share neural substrates with other cognitive processes that involve the MTL and MFC. Neurons in both areas prominently encode visual categories (*42*, *69*, *70*, *72*). We therefore started our analysis by examining the encoding of visual category in our dataset. Note that we restricted this analysis to the face and hand stimuli (scene stimuli were not examined for this analysis). As expected, neurons were modulated by visual category following stimulus onset (200 to 1200 ms) in both the MTL (65 of 370, 17.57%, binomial $P < 10^{-20}$; 42 neurons in AMY and 23 neurons in HIPP; see an example in Fig. 4A and group results in Fig. 4C) and the MFC (48 of 286, 16.61%, binomial $P = 9.70 \times 10^{-13}$; 20 neurons in dACC and 28 neurons in pre-SMA; see an example in Fig. 4B and group results in Fig. 4C). We refer to these neurons as category-selective (CS) neurons. Sixty-one of the 113 CS neurons (53.98%; see Fig. 4, D to F) showed higher activity for faces (face-preferring), with the remaining 52 (46.02%; Fig. 4, D, G, and H) showing a greater response to hands (hand-preferring). The proportions of the two types of neurons were comparable (Fig. 4D) in HIPP (face-preferring: 10 of 23, 43.48% versus

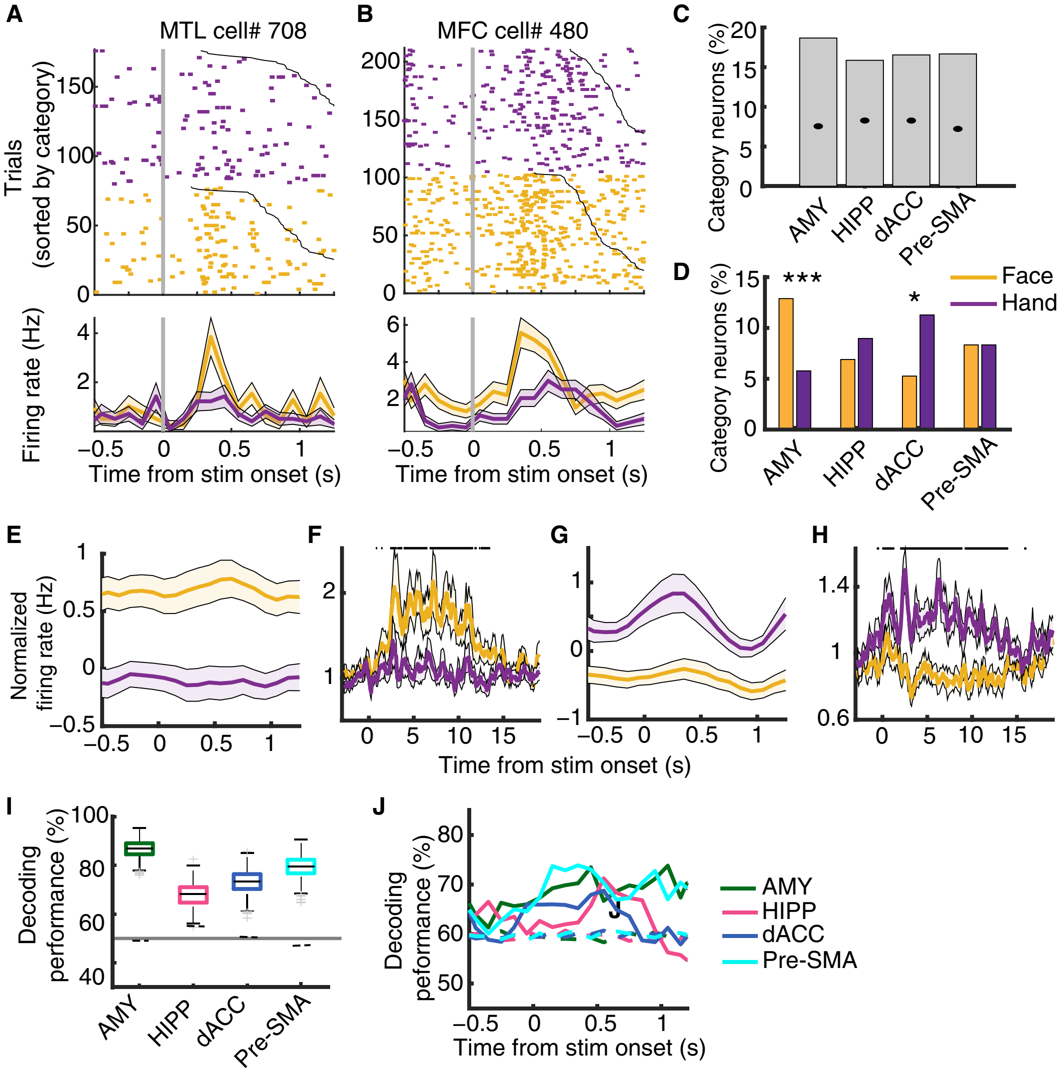

Fig. 4. Category selective neurons. (**A** and **B**) Two example neurons in the MTL (A) and MFC (B) discriminated between face and hand stimuli. Trials were grouped by category. (**C** and **D**) CS neuron percentages across the four recorded areas. (D) CS neurons that had a higher response to faces (yellow) and hands (purple) in each region. A χ^2 test of proportion was performed for CS percentages of face-preferring and hand-preferring neurons in each recorded area. ***$P < 0.001$ and *$P < 0.05$. (**E** to **H**) Average normalized firing rate for CS neurons. [(E) and (G)] Responses were aligned to the stimulus onset of each trial. [(F) and (H)] Responses were aligned to the stimulus onset of each block. [(E) and (F)] Face-preferring neurons (i.e., neurons had greater activity to face stimuli). [(G) and (H)] Hand-preferring neurons (i.e., neurons had greater activity to hand stimuli). (**I** and **J**) Population decoding of category with mean firing rate (I) and sliding time window (J), both on the whole population of each brain area. dACC, dorsal anterior cingulate cortex.

hand-preferring: 13 of 23, 56.52%) and pre-SMA (face-preferring: 15 of 30, 50% versus hand-preferring: 15 of 30, 50%). In contrast, the AMY had a higher proportion of face-preferring CS neurons (Fig. 4D; face-preferring: 29 of 42, 69.05%; hand-preferring: 13 of 42, 30.95%; χ^2 test of proportion: $P = 0.0005$) than hand-preferring CS neurons. Consistently, population decoding of category achieved an above-chance accuracy for each brain area (Fig. 4I; AMY: 86.74 ± 3.40% [mean ± SD], $P < 0.001$; HIPP: 68.01 ± 4.61% [mean ± SD], $P < 0.001$; dACC: 73.29 ± 4.57% [mean ± SD], $P < 0.001$; pre-SMA: 85.66 ± 4.36% [mean ± SD], $P < 0.001$). MFC neurons, including both dACC and pre-SMA, peaked earlier than the MTL neurons after stimulus onset, with HIPP peaked at the latest latency (Fig. 4J). In line with previous studies, our results confirmed that categorical information is prominently represented at both the individual neuron level and population level in both the MTL and MFC.

Specific representation of inference and category in the MTL and MFC

To examine whether the representation of stimulus categories was modulated by inference type, we first selected CS neurons (i.e., face versus hand) in why and how trials separately (Fig. 5, A and B). The proportion of CS neurons was similar in why (Fig. 5C**;** MTL: 46 of 370, percentage = 12.43%; MFC: 42 of 313, percentage = 13.42%) and how trials (MTL: 58 of 370, percentage = 15.68%; MFC: 46 of 313, percentage = 14.70%). Neurons selected during why trials were more likely to also be selected during how trials (Fig. 5D) in both the MTL (χ^2 test of the proportion of how-CS neurons among why-CS neurons versus all neurons: $P = 0.01$) and MFC ($P = 0.0002$). Similar results were revealed in subregions in the MTL and MFC (fig. S5, A and B). We confirmed this conclusion at the single-trial level: CS neurons selected during why trials

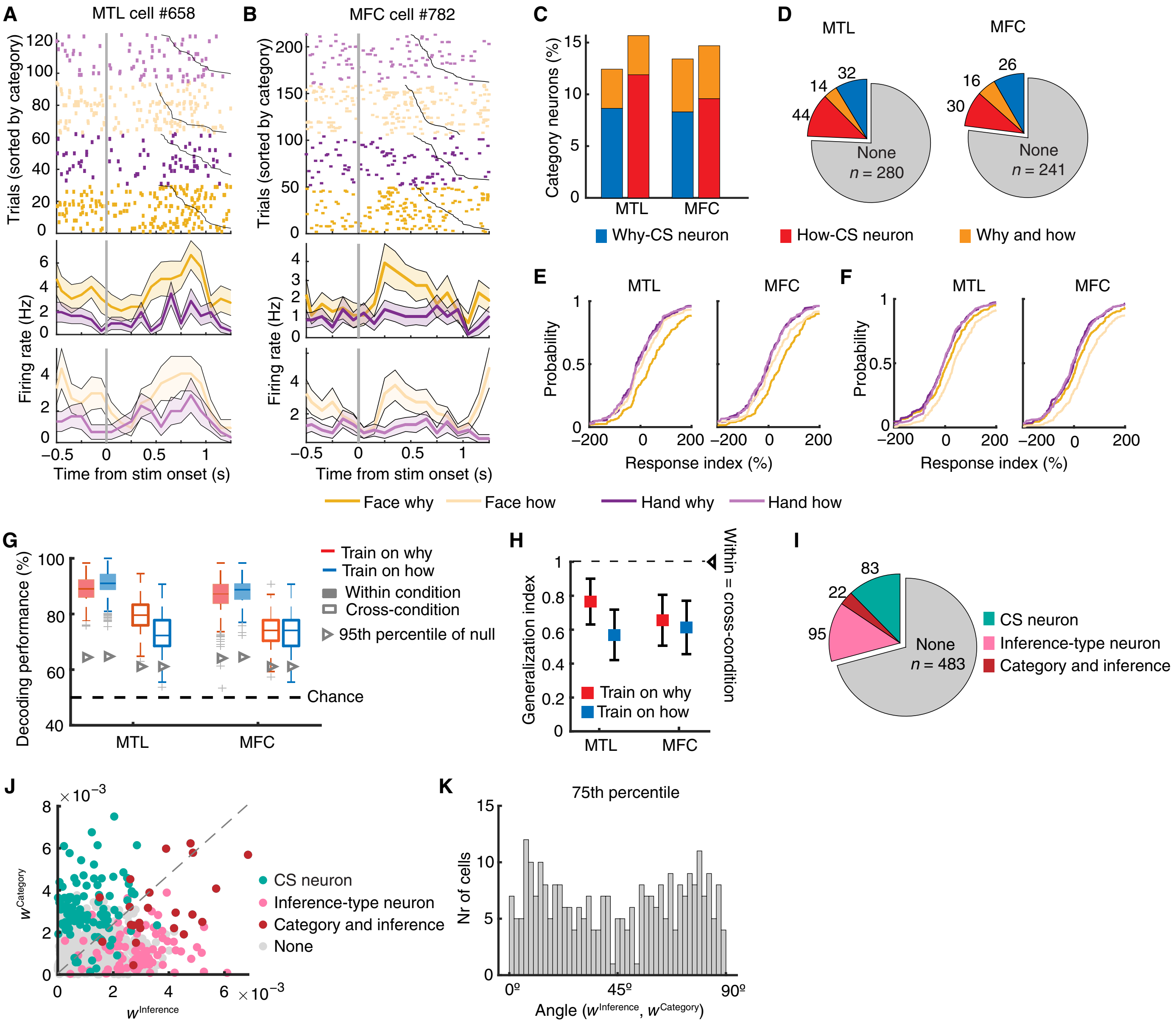

Fig. 5. Independent representation of inference type and visual category. (**A** and **B**) Example CS neurons that responded differently to face versus hand stimuli regardless of inference type. (**C**) Proportion of CS neurons in the MTL and MFC. (**D**) Distribution of CS neurons selected with why and how trials. (**E** and **F**) Single-trial RSI discriminating between categories in why-CS neurons (E) and how-CS neurons (F). The RSI for why and how is shown separately. Dark colors indicate why condition. Light colors indicate how condition. (**G**) Population decoding of category using within-inference decoder (i.e., train and test with why or how trials only) and cross-inference decoder (i.e., train with one inference type and test on the other type). (**H**) Generalization index for the cross-condition decoders quantified the generalizability of the cross-condition decoders of category. (**I**) Distribution of inference-type neurons and category neurons among the whole population. (**J** and **K**) Comparison of inference-type and CS cells using assigned decoder weight. (J) Scatter plot of the importance index (see Materials and Methods for details) assigned by a decoder to each cell in inference-type decoding (*x* axis) and category-decoding (*y* axis) of all cells. The features for the decoders are firing rates across the entire population in the MTL and MFC during the 0.2 to 1.2 s time window. Selected inference-type neurons and category neurons were superimposed on the plot with different colors (pink: inference; green: category; red: both). (K) Distribution of the weight vector angle for the top 25% of cells in either decoder (see Materials and Methods).

differentiated between faces versus hands also during how questions (Fig. 5E, light colors; KS test, MTL: $K = 0.08$, $P = 0.84 \times 10^{-5}$; MFC: $K = 0.10$, $P = 0.13 \times 10^{-7}$), and vice versa (Fig. 5F; how, MTL: $K = 0.24$, $P < 1 \times 10^{-56}$; MFC: $K = 0.26$, $P = 0.80 \times 10^{-55}$; why, MTL: $K = 0.09$, $P = 0.12 \times 10^{-8}$; MFC: $K = 0.11$, $P = 0.07 \times 10^{-8}$). These results suggested that inference type did not modulate the encoding of the category at the single-neuron level, indicating that the two variables are independent.

To confirm this impression at the population level, we next examined whether a decoder trained to distinguish between faces and hands when participants were making one type of inference would generalize to the other type of inference. Decoding performance was similar when making different types of inferences (Fig. 5E, filled bars; MTL: why: 88.90 ± 4.30% [mean ± SD] and how: 91.02 ± 3.97% [mean ± SD]; MFC: why: 87.00 ± 5.23% [mean ± SD] and how: 88.19 ± 4.24% [mean ± SD]). In line with the single-neuron level analysis, category decoding generalized between why and how (Fig. 5E; train with why test on how: 79.72 ± 4.91% [mean ± SD]; $P < 0.0001$ against null distribution; train with how test on why: 73.20 ± 5.85%) and MFC (train with why: 73.96 ± 4.99% [mean ± SD], $P < 0.0001$; train with how: 73.25 ± 5.65% [mean ± SD], $P < 0.0001$). More specifically, category decoding generalized well across inference types in the AMY, dACC, and pre-SMA, but not in the HIPP (fig. S5C). We further quantified the generalizability when decoding categories across inference types using the generalization index (Fig. 5H), which confirmed the above observations.

Above findings indicate that the representation of categories and inference relied on different sets of neurons. To test this hypothesis, we first examined the overlap between inference-type neurons and CS neurons. The two neural populations were largely distinct (Fig. 5I and see fig. S4, A and C for separate results in the MTL and MFC). This conclusion also held at the population level: Neurons that contributed strongly to decoding in one of the tasks tended to not strongly contribute to the other task and vice versa (Fig. 5, J and K and see fig. S4 B and E for results in the MTL and MFC separately; Fig. 5K; Hartigan dip test: dip = 0.10, $P < 0.0001$). We observed similar results in the MTL (fig. S4C) and MFC (fig. S4F) and a further breakdown in the four subregions (fig. S5, D and E) separately. Together, these results support the conclusion that the representation of inference and categorization was independent of each other, with visual category represented in a domain-general manner.

DISCUSSION

We investigated the neuronal mechanisms underlying social inference in the human MTL and MFC by examining how neural activity changed when participants made different types of inference. Consistent with previous neuroimaging findings, we observed neural representations of social inference type in the MFC. MTL neurons also encoded social inference type and more strongly so for faces. This finding is in contrast to neuroimaging studies that reveal no significant differences in the MTL for the why-how contrast (*36*, *49*, *61*). Furthermore, our results revealed a key difference between the MFC and MTL. In the MFC, representations of inference type were domain general across the different social categories (face versus hand), as shown by cross-condition generalization, but were domain specific when comparing social versus nonsocial domains (representations did not generalize across person versus natural scene). On the other hand, in the MTL, the neural representations were highly specialized (domain specific), with no generalization between faces, hands, or scenes. The striking entanglement of faces and hands, which are both social stimuli, with social inference in the MTL shows that, at least at this level of processing in the brain, there is not an abstract social inference computation. Rather, social inference is closely tied to the specific social stimulus category in the MTL. In contrast, the encoding of visual categories itself (i.e., selectivity for faces or hands) was domain general across task conditions in both brain areas. Together, our findings reveal that the MTL encodes inference type in a domain-specific manner in both social and nonsocial domains, indicating a fractionation of social inference processing that is tied to specific classes of social stimuli.

MTL is involved in social inference

Our finding that the MFC participates in social inference is consistent with earlier neuroimaging findings in the same task (*4*, *61*, *64*). Our contribution to this literature is that we reveal a single-neuron substrate of this common neuroimaging finding.

A contribution we make is that we identify neurons in the amygdala and hippocampus that fired differentially to different types of inference in both social and nonsocial domains. This is in contrast to neuroimaging work that shows no such differences at the fMRI-BOLD level (*49*, *61*). Our results support the hypothesis that the amygdala and hippocampus, as critical components of the social brain (*27*, *39*, *71*), are among the neurobiological bases of social inference.

Although only a few fMRI studies have linked the amygdala to social inference (*73*), the mixed results derived from studies of individuals with damaged amygdala (*49*–*51*) have led to debate about the role that the MTL plays in social inference. While Stone and colleagues (*50*) reported impaired performance of patients with bilateral amygdala lesions in two social inference tasks (“recognition of faux pas” and “reading the mind in the eyes”), a recent study suggested that “amygdala is not a necessary component of the cortical network for false-belief reasoning” (*49*). It is worth noting that in the latter study, the authors observed that amygdala activation in a false-belief task could only be revealed when a large number of participants were included. The discrepancies among these studies could potentially be explained by diversity in the tasks used and the extent of the lesion. Our study used a well-established task and found that the MTL exhibited a significant yet weaker representation compared to the MFC, providing evidence for the involvement of the MTL in social inference.

Functional specialization of the social inference network

The topic of domain specificity in social inference has been widely discussed given the diversity of stimulus categories and formats applied in past neuroimaging studies (*36*, *54*). One of the debates is whether the neural network commonly activated for social inference, including the MFC and TPJ, is used exclusively for social inference or subserves other functions also. An fMRI study conducted on nonhuman primates found that the pre-SMA and ACC were exclusively activated by social interactions with other agents but not for other stimulus conditions, including physical interaction, agents’ actions, and faces (*57*). However, fMRI studies on humans suggested that the MFC, which covers the pre-SMA and ACC as in the present study, plays a more general role in multiple neural functions (*54*, *59*, *60*). We found that the MFC represented inference type and category information in a domain general manner, supporting the hypothesis that the MFC is a central region for general processing of social information rather than specifically only for social inference.

Domain-specific inferences in the MTL and MFC

Did the inferences from our three types of stimuli (hands, faces, and scenes) use a common neural mechanism in the human brain (*4*, *64*)? It is plausible that the different types of stimuli require distinct neural substrates as social and nonsocial inference is usually implemented with input from distinguishable domains (people versus scenes). Domain-specific processing of faces, bodies, places, and objects has been proposed and indicated from numerous imaging studies (*66*, *68*, *74*), which might be one of the reasons for the putative neural dissociation of social versus nonsocial inference that we found. On the other hand, both social and nonsocial inference necessarily demand semantic memory to understand the contents of a visual scene and the possible hidden causes that explain what is shown in the stimulus (*64*, *75*). Therefore, making attributions of different types of events might be expected to recruit a domain-general process as well. This hypothesis was supported by previous meta-analyses conducted over imaging studies across various social inference tasks (*23*). A recent fMRI study also showed that the majority of the brain regions identified in social attributions were also activated by nonsocial attributions (*64*).

The present study shows evidence for domain-specific processing of inference information across different categories of stimuli (face, hand, and scene) in the MTL. In contrast, the inference representation in the MFC (and, particularly, in the pre-SMA) was partially domain general. Social inference type was domain general across different social stimuli (faces and hands), but not when comparing social versus nonsocial stimuli. This finding is compatible with the interpretation that the processes in MTL remain domain specific because it reads out information of faces, hands, and scenes from different specific areas in the higher visual cortex that are also domain specific (*76*). In contrast, the MFC is a central region that plays an executive role in the use of world knowledge and domain-specific information fed from the MTL. Hence, among the areas we examined, only the MTL was found to contain specialized processes for social inference from faces and hands.

Our results can also be understood from the point of view of population-level coding. The absence of cross-condition generalization for inference type decoding between faces and hands indicates that the two variables (inference type and face/hand) are not represented independently in the MTL (Fig. 3) (*46*). Rather, they are entangled (*77*). Category decoding had high cross-condition generalization in the MTL (Fig. 5), indicating that it is specifically the inference type variable that was entangled with stimulus type.

Caveats and future direction

The current study has a number of caveats. First, we focused on inference processes with participants engaged as observers only. While sometimes overlapping, observing and performing actions might rely on different neural substrates (*78*). Real interactions might rely on different neural mechanisms (*79*, *80*). Second, we provided explicit instructions. Spontaneous social inferences, which are common (*81*), might rely on different mechanisms. These caveats might be addressed by using interactive paradigms, in which the participants act as agents and interact with other agents. Third, the frontal recordings we examined were exclusively from two locations along the medial wall. It remains an open question how related findings made in the lateral frontal areas (*82*) and other parts of the medial frontal wall that we did not sample. Fourth, we did not examine other parts of the MTL, particularly the parahippocampal gyrus and entorhinal cortex.

MATERIALS AND METHODS

Patients

We collected data in 19 sessions from 14 adult surgical epilepsy patients (five males) (table S1). All participants provided written informed consent using protocols approved by the Institutional Review Board of Cedars-Sinai Medical Center and the California Institute of Technology.

Task and procedure

In this study, we used a why/how task established in previous fMRI studies (*61*, *64*). A sequence of naturalistic pictures was presented to the patients using a block design. The task required patients to make social inference about the intention (why the person does this) or perceptual judgment about the action (how the person does this) of the person displayed in the pictures. Hand and face stimuli were evenly distributed across blocks ($n = 8$) and paired with one why or how question (see table S2 for the block questions, $n = 16$ in total). Each block contained eight different images from one category, resulting in 32 trials for each category in each condition. A pre-block cue (e.g., "is this person admiring someone?") was displayed for 2.1 s at the beginning of each block. After each trial, a brief verbal cue (e.g., "admiring?") was displayed for 0.15 s (Fig. 1A). Participants had a maximum of 1.7 s to respond after the stimulus onset. The task advanced immediately once the answer was given, or the display duration of the current stimulus reached 1.7 s. Each question was paired with five pictures designed to elicit a yes response, and three pictures designed to elicit a no response. No feedback was provided. The task thus featured a two (categories: face/hand) by two (questions: why/how) by two (choices: "yes/no") design, resulting in a total of 128 trials. The onset of each block was fixed across subjects and was designed to maximize the efficiency of separately estimating the contrast of interest (why versus how) for each of the two image categories. A varied length of interblock interval was thus implemented at the end of each block to keep the block onset synchronized across subjects. Besides, the order of why and how blocks were counterbalanced within each image category.

A subset of patients ($n = 8$; 9 sessions) completed an updated version of the why/how task (*64*), which included scene photographs depicting natural events in addition to a new set of face and hand images. In this version of the task, there were six blocks for each image category, resulting in a total of 36 blocks also with paired questions (see table S3 for the question list). Each block contained nine different images from one of the three categories, resulting in 54 trials for each category for why and how, respectively. Each question was paired with five pictures designed to elicit a yes response, and four pictures designed to elicit a no response. Stimuli were presented using MATLAB with the Psychtoolbox 3 (http://psychtoolbox.org).

Electrophysiology

We recorded from the bilateral amygdala (AMY), hippocampus (HIPP), dACC, and pre-SMA using implanted hybrid depth electrodes with eight macro contacts and eight microwires (see Fig. 1B for recording locations). The target locations in these recording sites were determined based solely on clinical need by the neurosurgeon and verified using post-implantation computed tomography and magnetic resonance imaging. We recorded continuous extracellular signals in the broadband range of 0.1 to 9000 Hz with a sampling rate of 32 kHz (ATLAS System, Neuralynx Inc.).

Spike sorting and single-neuron analysis

The raw signal underwent filtering with a zero-phase lag filter within the 300- to 3000-Hz band. Spike detection and sorting were performed using a semi-automated template-matching algorithm (*83*).

All peri-stimulus time histograms were computed using a 500-ms window with a step size of 100 ms, and no smoothing was applied. Neurons with an average firing rate greater than 0.2 Hz throughout the entire trial were kept for analysis.

Electrode localization

Electrode localization was achieved through post-operative MRI scans, which were registered to pre-operative MRI scans using Freesurfer's mri_robust_register (*70*, *84*). This procedure ensured accurate and subject-specific localization. To allow for comparability across studies, we summarized the electrode positions across subjects by aligning the locations to the MNI152-aligned CIT168 template brain (*85*) using an affine transformation followed by a symmetric image normalization (SyN) diffeomorphic transform (*86*). This method provided MNI coordinates for each recording location in this study (Fig. 1B).

Selection of inference-type and CS neurons using ANOVA model

Only trials with a response (96.38 ± 5.90% [mean ± SD]) were included in further analysis. To identify neurons that discriminated inference types or stimulus categories while taking into account all possible contributing factors, we applied a three-way ANOVA on each neuron. In the ANOVA model, we labeled each trial for different stimulus categories (face/hand), inference type (why/how), and choice (yes/no). A neuron that had a significant main effect ($P < 0.05$) on inference was defined as an inference-type neuron. Similarly, a neuron that had a significant main effect on category was identified as a CS neuron. A binomial test was conducted to determine the significance of the number of the selected neurons. A null distribution was created to further validate the significance by randomly reshuffling the inference or category label 1000 times and repeating the above selection procedure.

Selection of inference-type and CS neurons using *t* test

To check the modulation of one variable to another (i.e., category to inference type), we used an unpaired *t*-test to select neurons with one type of stimuli. For example, we selected inference-type neurons using face and hand stimuli separately.

Selection of inference-type and CS neurons using a linear mixed-effect model

To account for the effect of response time (RT), we selected the inference-type and CS neurons using a linear mixed-effect model [firing rate ~ category + inference + choice + (1|RT)], where RT was considered as a random factor. A neuron was considered to be an inference-type neuron if the fixed effect of inference passed $P < 0.05$. Similarly, a neuron was considered to be a CS neuron if the fixed effect of category passed $P < 0.05$. Neurons selected with this procedure were then compared with that selected with ANOVA.

Population decoding

Single-trial population decoding was conducted on pseudo-random population assembled across sessions to substitute for simultaneous recordings (*87*). We performed linear support vector machine (SVM) decoding on two contrasts: (i) inference type: why versus how, and (ii) category: face versus hand. As 22 more trials were displayed for each condition (e.g., face images with why questions) in the updated-version task (number of trials: $n = 54$; see the "Task and procedure" section), for these sessions, we split the trials in half and estimated the mean between them to generate a similar number of trials ($n = 27$) as the other sessions ($n = 32$). To further match the number of trials between different sessions, we excluded the first trial of each block and the last trial of the last block, resulting in 27 trials in each condition for the original version of the task (10 sessions). We randomly selected 75% of the whole neuron population in each interested brain area on each iteration of the decoder. The procedure was iterated for 500 times. To test the significance of the decoding performance, a null distribution was estimated by shuffling the labels of the conditions in each iteration. We then compared the average performance of the observed decoding performance with the null distribution (the *P* value was estimated by the rate of the null decoding exceeding the observed decoding). To compare the performance between different decoders, we constructed an empirical null distribution using paired differences of performances obtained with shuffled labels. The significance of the difference in performance between the two decoders was then determined by comparing the observed difference against the null difference distribution.

Upon feeding into the decoder, the data was first baseline corrected using interblock interval and then normalized (*z* scored) to account for any drift in the baseline period and the scale problem. A 10-fold cross-validation procedure was then performed to estimate the decoding accuracy for each contrast. The analysis was done in MATLAB by implementing the function "fitcsvm" with a kernel scale equal to 1. Decoding accuracy was displayed either as a function of time or in a fixed time window. Time course decoding was performed on the firing rate calculated in a 500-ms sliding window, with a step of 100 ms. For fixed-window decoding, we used the firing rate estimated in a time window from 200 to 1200 ms after stimulus onset.

To perform within-condition decoding (e.g., within-category decoding of inference), we trained and tested the decoder with trials of one condition only. We used the same procedure as described above.

To perform cross-condition decoding (e.g., cross-category decoding of inference), we trained the decoder with trials of one condition and tested with trials of the other condition. We used the same procedure as described above except that we only split the training set and testing set in half rather than 10-folds.

Single-trial RSI

For each neuron, we quantified whether its response differed between contrasted conditions using a single-trial RSI (see Eqs. 1 and 2), which has been proven to be effective in previous single-neuron studies (*88*, *89*). We reported this measurement for both inference-type and CS neurons, where we contrasted between why versus how and face versus hand trials. Typically, the RSI facilitates group analysis and comparisons between different types of cells (i.e., social inference preferring versus perceptual judgment preferring cells in this study). The RSI quantifies the response during why trials relative to the mean response during how trials and baseline. The mean response and baseline were calculated individually for each neuron

$$\text{Cell type 1: RSI}_i = \frac{\text{FR}_i - \text{mean}\left(\text{FR}_{\text{type1}}\right)}{\text{mean}\left(\text{FR}_{\text{baseline}}\right)} \cdot 100 \qquad (1)$$

$$\text{Cell type 2: RSI}_i = -\frac{\text{FR}_i - \text{mean}\left(\text{FR}_{type1}\right)}{\text{mean}\left(\text{FR}_{\text{baseline}}\right)} \cdot 100 \qquad (2)$$

For each trial, i, RSI_i is the baseline normalized mean firing rate (FR) in a fixed time window from 200 to 1200 ms after stimulus onset (the same time interval as cell selection). The baseline is the mean firing rate estimated within the first 2 s before the onset of the first trial in each block.

The cumulative distribution function (CDF) was constructed by calculating for each possible value x of the RSI by counting how many examples are smaller than x. That is, $F(x) = P\ (\mathbf{X} \leq x)$, where $\mathbf{X}$ is a vector of all RSI values. The CDFs of different conditions (why versus how; face versus hand) were compared using two-tailed two-sample KS tests.

Generalization index

We defined a generalization index (Eq. 3) to compare the within-condition decoding to the across condition generalization (*70*)

$$g = \frac{\text{Cross} - \text{Chance}}{\text{Within} - \text{Chance}} \qquad (3)$$

where "Within" indicates the decoding performance for within-condition decoders (e.g., train on face and test on face), "Cross" indicates the decoding performance for cross-condition decoders (e.g., train on face and test on hand), and "Chance" indicates the theoretical chance level of decoding performance for the variable of interest (inference = 0.5, category = 0.5).

Normalized weight metric

$$w_i^t = \frac{|\,w_i^t\,|}{\sum_i^n |\,w_i^t\,|}$$

We further determined the extent of specialization of each neuron using the angle between the vector of $(w^{\text{Inference}}, w^{\text{Category}})$ with respect to the x axis.

Statistical significance

Statistical significance for all tests was set at $P < 0.05$. We corrected for multiple comparisons over time points by the false discovery rate (see the description in Figs. 2G and 4G) method (*90*). For all other tests, we adjusted the threshold of the P value for multiple comparisons, where appropriate, using the Holm-Bonferroni correction (*91*) to control the family-wise error rate.

Supplementary Materials

This PDF file includes:
Supplementary Text
Tables S1 to S3
Figs. S1 to S5

REFERENCES AND NOTES

1. D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**, 515–526 (1978).
2. J. Call, M. Tomasello, Does the chimpanzee have a theory of mind? 30 years later. *Trends Cogn. Sci.* **12**, 187–192 (2008).
3. S. Ondobaka, J. Kilner, K. Friston, The role of interoceptive inference in theory of mind. *Brain Cogn.* **112**, 64–68 (2017).
4. R. P. Spunt, R. Adolphs, The neuroscience of understanding the emotions of others. *Neurosci. Lett.* **693**, 44–48 (2019).
5. H. M. Wellman, D. Cross, J. Watson, Meta-analysis of theory-of-mind development: The truth about false belief. *Child Dev.* **72**, 655–684 (2001).
6. C. D. Frith, U. Frith, The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
7. S. Baron-Cohen, A. M. Leslie, U. Frith, Does the autistic child have a "theory of mind"? *Cognition* **21**, 37–46 (1985).
8. U. Frith, F. Happé, Autism: Beyond "theory of mind". *Cognition* **50**, 115–132 (1994).
9. R. Corcoran, G. Mercer, C. D. Frith, Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophr. Res.* **17**, 5–13 (1995).
10. M. Alegre, J. Guridi, J. Artieda, The mirror system, theory of mind and Parkinson's disease. *J. Neurol. Sci.* **310**, 194–196 (2011).
11. M. Freedman, D. T. Stuss, Theory of mind in Parkinson's disease. *J. Neurol. Sci.* **310**, 225–227 (2011).
12. D. P. Kennedy, R. Adolphs, The social brain in psychiatric and neurological disorders. *Trends Cogn. Sci.* **16**, 559–572 (2012).
13. A. Gopnik, L. Schulz, Mechanisms of theory formation in young children. *Trends Cogn. Sci.* **8**, 371–377 (2004).
14. K. N. Ochsner, M. D. Lieberman, The emergence of social cognitive neuroscience. *Am. Psychol.* **56**, 717–734 (2001).
15. T. Rusch, S. Steixner-Kumar, P. Doshi, M. Spezio, J. Glascher, Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia* **146**, 107488 (2020).
16. D. C. Dennett, Intentional Systems. *J. Philos.* **68**, 87–106 (1971).
17. H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**, 103–128 (1983).
18. S. Baron-Cohen, M. O'Riordan, V. Stone, R. Jones, K. Plaisted, Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *J. Autism Dev. Disord.* **29**, 407–418 (1999).
19. I. A. Apperly, F. Warren, B. J. Andrews, J. Grant, S. Todd, Developmental continuity in theory of mind: Speed and accuracy of belief-desire reasoning in children and adults. *Child Dev.* **82**, 1691–1703 (2011).
20. G. Ganis, S. M. Kosslyn, S. Stose, W. L. Thompson, D. A. Yurgelun-Todd, Neural correlates of different types of deception: An fMRI investigation. *Cereb. Cortex* **13**, 830–836 (2003).
21. M. I. Gobbini, A. C. Koralek, R. E. Bryan, K. J. Montgomery, J. V. Haxby, Two takes on the social brain: A comparison of theory of mind tasks. *J. Cogn. Neurosci.* **19**, 1803–1814 (2007).
22. H. L. Gallagher, C. D. Frith, Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* **7**, 77–83 (2003).
23. 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).
24. 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).
25. M. Aichhorn, J. Perner, B. Weiss, M. Kronbichler, W. Staffen, G. Ladurner, Temporo-parietal junction activity in theory-of-mind tasks: Falseness, beliefs, or attention. *J. Cogn. Neurosci.* **21**, 1179–1192 (2009).
26. L. Brothers, The social brain : A project for integrating primate behavior and neuropsychology in a new domain. *Concepts Neurosci.* **1**, 27–51 (1990).
27. R. Adolphs, The social brain: Neural basis of social knowledge. *Annu. Rev. Psychol.* **60**, 693–716 (2009).
28. Z. Fu, A. Sajad, S. P. Errington, J. D. Schall, U. Rutishauser, Neurophysiological mechanisms of error monitoring in human and non-human primates. *Nat. Rev. Neurosci.* **24**, 153–172 (2023).
29. A. de la Vega, L. J. Chang, M. T. Banich, T. D. Wager, T. Yarkoni, Large-scale meta-analysis of human medial frontal cortex reveals tripartite functional organization. *J. Neurosci.* **36**, 6553–6562 (2016).
30. P. Silva Moreira, P. Marques, R. Magalhaes, Identifying functional subdivisions in the medial frontal cortex. *J. Neurosci.* **36**, 11168–11170 (2016).
31. P. C. Fletcher, F. Happe, U. Frith, S. C. Baker, R. J. Dolan, R. S. J. Frackowiak, C. D. Frith, Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition* **57**, 109–128 (1995).
32. H. L. Gallagher, F. Happe, N. Brunswick, P. C. Fletcher, U. Frith, C. D. Frith, Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* **38**, 11–21 (2000).
33. E. Brunet, Y. Sarfati, M.-C. Hardy-Bayle, J. Decety, A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage* **11**, 157–166 (2000).

34. B. A. Vollm, A. N. W. Taylor, P. Richardson, R. Corcoran, J. Stirling, S. McKie, J. F. W. Deakin, R. Elliott, Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. *NeuroImage* **29**, 90–98 (2006).
35. T. P. German, J. L. Niehaus, M. P. Roarty, B. Giesbrecht, M. B. Miller, Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *J. Cogn. Neurosci.* **16**, 1805–1817 (2004).
36. S. J. Carrington, A. J. Bailey, Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Hum. Brain Mapp.* **30**, 2313–2335 (2009).
37. A. D. Rowe, P. R. Bullock, C. E. Polkey, R. G. Morris, 'Theory of mind' impairments and their relationship to executive functioning following frontal lobe excisions. *Brain* **124**, 600–616 (2001).
38. C. M. Bird, F. Castelli, O. Malik, U. Frith, M. Husain, The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain* **127**, 914–928 (2004).
39. A. Montagrin, C. Saiote, D. Schiller, The social hippocampus. *Hippocampus* **28**, 672–679 (2018).
40. L. R. Squire, C. E. Stark, R. E. Clark, The medial temporal lobe. *Annu. Rev. Neurosci.* **27**, 279–306 (2004).
41. R. Cao, C. Lin, J. Hodge, X. Li, A. Todorov, N. J. Brandmeir, S. Wang, A neuronal social trait space for first impressions in the human amygdala and hippocampus. *Mol. Psychiatry* **27**, 3501–3509 (2022).
42. G. Kreiman, C. Koch, I. Fried, Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* **3**, 946–953 (2000).
43. U. Rutishauser, O. Tudusciuc, S. Wang, A. N. Mamelak, I. B. Ross, R. Adolphs, Single-neuron correlates of atypical face processing in autism. *Neuron* **80**, 887–899 (2013).
44. S. Wang, O. Tudusciuc, A. N. Mamelak, I. B. Ross, R. Adolphs, U. Rutishauser, Neurons in the human amygdala selective for perceived emotion. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3110–E3119 (2014).
45. A. O. Constantinescu, J. X. O'Reilly, T. E. J. Behrens, Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
46. H. S. Courellis, J. Minxha, A. R. Cardenas, D. Kimmel, C. M. Reed, T. A. Valiante, C. D. Salzman, A. N. Mamelak, S. Fusi, U. Rutishauser, Abstract representations emerge in human hippocampal neurons during inference behavior. bioRxiv 566490 [Preprint] (2023). https://doi.org/10.1101/2023.11.10.566490.
47. J. S. Simons, H. J. Spiers, Prefrontal and medial temporal lobe interactions in long-term memory. *Nat. Rev. Neurosci.* **4**, 637–648 (2003).
48. J. A. Gordon, Oscillations and hippocampal-prefrontal synchrony. *Curr. Opin. Neurobiol.* **21**, 486–491 (2011).
49. R. P. Spunt, J. T. Elison, N. Dufour, R. Hurlemann, R. Saxe, R. Adolphs, Amygdala lesions do not compromise the cortical network for false-belief reasoning. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4827–4832 (2015).
50. V. E. Stone, S. Baron-Cohen, A. Calder, J. Keane, A. Young, Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia* **41**, 209–220 (2003).
51. P. Shaw, E. J. Lawrence, C. Radbourne, J. Bramham, C. E. Polkey, A. S. David, The impact of early and late damage to the human amygdala on 'theory of mind' reasoning. *Brain* **127**, 1535–1548 (2004).
52. U. Rutishauser, O. Tudusciuc, D. Neumann, A. N. Mamelak, A. C. Heller, I. B. Ross, L. Philpott, W. W. Sutherling, R. Adolphs, Single-unit responses selective for whole faces in the human amygdala. *Curr. Biol.* **21**, 1654–1660 (2011).
53. S. Wang, R. J. Yu, J. M. Tyszka, S. S. Zhen, C. Kovach, S. Sun, Y. Huang, R. Hurlemann, I. B. Ross, J. M. Chung, A. N. Mamelak, R. Adolphs, U. Rutishauser, The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nat. Commun.* **8**, 14821 (2017).
54. R. P. Spunt, R. Adolphs, A new look at domain specificity: Insights from social neuroscience. *Nat. Rev. Neurosci.* **18**, 559–567 (2017).
55. D. Kliemann, R. Adolphs, The social neuroscience of mentalizing: Challenges and recommendations. *Curr. Opin. Psychol.* **24**, 1–6 (2018).
56. N. Kanwisher, Domain specificity in face perception. *Nat. Neurosci.* **3**, 759–763 (2000).
57. J. Sliwa, W. A. Freiwald, A dedicated network for social interaction processing in the primate brain. *Science* **356**, 745–749 (2017).
58. P. H. Rudebeck, M. J. Buckley, M. E. Walton, M. F. S. Rushworth, A role for the macaque anterior cingulate gyrus in social valuation. *Science* **313**, 1310–1312 (2006).
59. K. R. Ridderinkhof, S. Nieuwenhuis, T. S. Braver, Medial frontal cortex function: An introduction and overview. *Cogn. Affect Behav. Neurosci.* **7**, 261–265 (2007).
60. P. A. Kragel, M. Kano, L. Van Oudenhove, H. G. Ly, P. Dupont, A. Rubio, C. Delon-Martin, B. L. Bonaz, S. B. Manuck, P. J. Gianaros, M. Ceko, E. A. Reynolds Losin, C.-W. Woo, T. E. Nichols, T. D. Wager, Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* **21**, 283–289 (2018).
61. R. P. Spunt, R. Adolphs, Validating the why/how contrast for functional MRI studies of theory of mind. *NeuroImage* **99**, 301–311 (2014).
62. L. S. King, V. C. Salo, A. Kujawa, K. L. Humphreys, Advancing the RDoC initiative through the assessment of caregiver social processes. *Dev. Psychopathol.* **33**, 1648–1664 (2021).
63. R. P. Lobo, K. L. Bottenhorn, M. C. Riedel, A. I. Toma, M. M. Hare, D. D. Smith, A. C. Moor, I. K. Cowan, J. A. Valdes, J. E. Bartley, T. Salo, E. R. Boeving, B. Pankey, M. T. Sutherland, E. D. Musser, A. R. Laird, Neural systems underlying RDoC social constructs: An activation likelihood estimation meta-analysis. *Neurosci. Biobehav. Rev.* **144**, 104971 (2023).
64. R. P. Spunt, R. Adolphs, Folk explanations of behavior: A specialized use of a domain-general mechanism. *Psychol. Sci.* **26**, 724–736 (2015).
65. A. Tusche, R. Spunt, L. Paul, J. Tyszka, R. Adolphs, Neural signatures of social inferences predict the number of real-life social contacts and autism severity. *Nat. Commun.* **14**, 4399 (2023).
66. N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
67. N. Kanwisher, G. Yovel, The fusiform face area: A cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2109–2128 (2006).
68. R. F. Schwarzlose, J. D. Swisher, S. Dang, N. Kanwisher, The distribution of category and location information across object-selective regions in human visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4447–4452 (2008).
69. T. P. Reber, M. Bausch, S. Mackay, J. Bostrom, C. E. Elger, F. Mormann, Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe. *PLOS Biol.* **17**, e3000290 (2019).
70. J. Minxha, R. Adolphs, S. Fusi, A. N. Mamelak, U. Rutishauser, Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**, eaba3313 (2020).
71. R. Adolphs, D. Tranel, A. R. Damasio, The human amygdala in social judgment. *Nature* **393**, 470–474 (1998).
72. U. Rutishauser, L. Reddy, F. Mormann, J. Sarnthein, The architecture of human memory: Insights from human single-neuron recordings. *J. Neurosci.* **41**, 883–890 (2021).
73. S. Baron-Cohen, H. A. Ring, S. Wheelwright, E. T. Bullmore, M. J. Brammer, A. Simmons, S. C. R. Williams, Social intelligence in the normal and autistic brain: An fMRI study. *Eur. J. Neurosci.* **11**, 1891–1898 (1999).
74. N. Kanwisher, The quest for the FFA and where it led. *J. Neurosci.* **37**, 1056–1061 (2017).
75. J. R. Binder, R. H. Desai, W. W. Graves, L. L. Conant, Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
76. F. D. Raslau, I. T. Mark, A. P. Klein, J. L. Ulmer, V. Mathews, L. P. Mark, Memory part 2: The role of the medial temporal lobe. *Am. J. Neuroradiol.* **36**, 846–849 (2015).
77. I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, A. Lerchner, Towards a definition of disentangled representations. arXiv:1812.02230 [cs.LG] (2018).
78. R. Mukamel, A. D. Ekstrom, J. Kaplan, M. Iacoboni, I. Fried, Single-neuron responses in humans during execution and observation of actions. *Curr. Biol.* **20**, 750–756 (2010).
79. L. Schilbach, B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, K. Vogeley, Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414 (2013).
80. H. Z. G. Probolovski, Commentary: Using second-person neuroscience to elucidate the mechanisms of reciprocal social interaction. *Front. Behav. Neurosci.* **14**, 13 (2020).
81. S. Baek, M. Song, J. Jang, G. Kim, S.-B. Paik, Spontaneous generation of face recognition in untrained deep neural networks. bioRxiv 857466 [Preprint] (2019). https://doi.org/10.1101/857466.
82. M. Jamali, B. L. Grannan, E. Fedorenko, R. Saxe, R. Baez-Mendoza, Z. M. Williams, Single-neuronal predictions of others' beliefs in humans. *Nature* **591**, 610–614 (2021).
83. U. Rutishauser, E. M. Schuman, A. N. Mamelak, Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci. Methods* **154**, 204–224 (2006).
84. M. Reuter, H. D. Rosas, B. Fischl, Highly accurate inverse consistent registration: A robust approach. *NeuroImage* **53**, 1181–1196 (2010).
85. J. M. Tyszka, W. M. Pauli, In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. *Hum. Brain Mapp.* **37**, 3979–3998 (2016).
86. B. Avants, J. T. Duda, J. Kim, H. Zhang, J. Pluta, J. C. Gee, J. Whyte, Multivariate analysis of structural and diffusion imaging in traumatic brain injury. *Acad. Radiol.* **15**, 1360–1375 (2008).
87. E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, T. Poggio, Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.* **100**, 1407–1419 (2008).
88. S. Wang, A. N. Mamelak, R. Adolphs, U. Rutishauser, Encoding of target detection during visual search by single neurons in the human brain. *Curr. Biol.* **28**, 2058–2069.e4 (2018).
89. R. Cao, X. Li, N. J. Brandmeir, S. Wang, Encoding of facial features by single neurons in the human amygdala and hippocampus. *Commun. Biol.* **4**, 1394 (2021).
90. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
91. H. G. Mikel Aickin, Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *Am. J. Public Health* **86**, 726–728 (1996).

Acknowledgments: We thank all patients for their participation, staff from Cedars-Sinai Medical Center for support with patient testing, and H. Courellis for discussion. **Funding:** This research was supported by the McDonnell Center for Systems Neuroscience, AFOSR (FA9550-21-1-0088), NSF (BCS-1945230 and IIS-2114644), and NIH (K99EY036650 to R.C., R01MH129426 to S.W., P50MH094258 to R.A., U01NS117839 to U.R., and R01MH134990 to U.R. and R.A.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Author contributions:** J.D., R.A., and U.R. designed the research. J.D. and U.R. performed the experiments. A.N.M. performed the surgery. R.C., S.W., and U.R. analyzed the data. R.C., R.A., S.W., and U.R. wrote the paper. All authors discussed the results and contributed toward the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in this paper are present in the paper and/or the Supplementary Materials. The raw data (single-neuron activity and behavior) is available at https://osf.io/qx92s/.

Submitted 12 February 2024
Accepted 1 November 2024
Published 4 December 2024
10.1126/sciadv.ado6166