Published in final edited form as:
Nature. 2021 March ; 591(7851): 610–614. doi:10.1038/s41586-021-03184-0.

Single-neuronal predictions of others' beliefs in humans

Mohsen Jamali[1], **Benjamin L. Grannan**[1], **Evelina Fedorenko**[2], **Rebecca Saxe**[2], **Raymundo Báez-Mendoza**[1], **Ziv M. Williams**[1,3,4,*]

[1]Department of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston MA

[2]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston MA

[3]Harvard-MIT Division of Health Sciences and Technology, Boston MA

[4]Program in Neuroscience, Harvard Medical School, Boston MA

Abstract

Human social behavior crucially depends on our ability to reason about others. This capacity for 'theory of mind' plays a vital role in social cognition because it allows us not only to form a detailed understanding of the hidden thoughts and beliefs of other individuals but to also understand that they may differ from our own[1–3]. Although a number of areas in the human brain have been linked to social reasoning[4, 5] and its disruption across a variety of psychosocial disorders[6–8], the basic cellular mechanisms that underlie human theory of mind remain undefined. Using a rare opportunity to acutely record from single cells in the human dorsomedial prefrontal cortex, we discover neurons that reliably encode information about others' beliefs across richly varying scenarios and that distinguish self- from other-belief related representations. By further following their encoding dynamics, we show how these cells represent the contents of the other's beliefs and accurately predict whether they are true or false. We also show how they track inferred beliefs from another's specific perspective and how their activities relate to behavioral performance. Together, these findings reveal a detailed cellular process in the human dorsomedial prefrontal cortex for representing another's beliefs and identify candidate neurons that could support theory of mind.

Humans have the ability to form remarkably detailed representations about other individuals and to understand that others may hold thoughts or beliefs that are distinct from their own[1, 3, 9]. This capacity for 'theory of mind' develops early during human ontogeny[3, 10, 11] and plays a vital role in social cognition. Yet, unlike most sensorimotor processes that are based on the observed relation between sensory input, actions and outcome and which have

[*]corresponding author, zwilliams@mgh.harvard.edu.

Contributions: M.J performed the analysis, M.J, B.G. and R.B.-M performed the experiments, E.F. and R.S. provided feedback and Z.M.W conceived and designed the study, performed the experiments, obtained the recordings and oversaw the project.

Competing interests: The Authors declare no competing financial or non-financial interests.

Code availability: All software used in this study are listed in the Reporting Summary along with their versions. The primary MATLAB codes used to perform the statistical and data analyses in this study are available from the corresponding author upon reasonable request.

been broadly studied in animal models [12], little is known about the single-neuronal mechanisms that underlie theory of mind.

Functional imaging studies have provided an important understanding of the network of brain areas that supports social reasoning including the temporal-parietal junction (TPJ), parts of the superior temporal sulcus and dorsal medial prefrontal cortex (dmPFC)[4, 13, 14]. The TPJ, for example, has been shown to display changes in activity when individuals form mental representations of others or their beliefs[4, 5], and the dmPFC has been found to activate when attributing mental states to others[3, 15–17] or when distinguishing another's beliefs from reality[4, 5, 10, 17, 18]. The precise cellular constructs and logic by which humans reason about others or represent their beliefs, however, remain largely unknown.

A critical test for theory of mind is the false belief task which requires individuals to make inferences about another's beliefs[1, 4, 5, 19]. In these tasks, a participant may be given a brief story narrative describing a social agent that may or may not hold false beliefs about events in their worlds[4, 17, 20]. For example, they may be given a narrative such as "You and Tom see a jar on the table. After Tom leaves, you move the jar to the cupboard", followed by the question "Where does Tom think the jar is?". This approach, therefore, incorporates two core components thought to be essential for theory of mind – the ability to reason about the beliefs of other individuals and the ability to distinguish another's beliefs and perspective of reality from an individual's own. Here, we took the opportunity to record from single neurons in the superior frontal gyrus of the human dmPFC – an area previously implicated in social reasoning and theory of mind[3, 15–18] – in order to begin investigating these processes at the cellular level.

Neuronal predictions of another's beliefs

We used custom-adapted multi-electrode microarrays (Fig. 1a) to stably record from 212 well-isolated single units in the dmPFC (Extended Data Fig. 1) of 11 participants (Extended Data Table 1a) as they performed a verbal version of the false belief task[21]. Another 112 single units were recorded from 4 participants performing additional controls, for a total of 324 putative neurons. All trial events were aligned to neural activity at millisecond resolution and analyzed off-line (Fig. 1b, Methods).

To first distinguish neuronal signals that reflect another's beliefs from those that may reflect other more generalized non-social representations[17, 22], the participants were given brief story narratives followed by questions about them (Extended Data Table 2a). Here, the narratives provided richly detailed information about social agents and events in their worlds, requiring the participant to consider either another's beliefs of reality (i.e., *other-belief* trials) or its physical state (i.e., *physical* trials; Fig. 1c, 1e *left*). Whereas both trial conditions entailed a discrepancy between the past and present state of reality (e.g., as the result of moving a jar from a table to a cupboard), only the former required the participant to consider another's beliefs. To further identify neural signals that may reflect another's specific beliefs rather than simply any beliefs, we also required the participants to consider others' beliefs that were either distinctly false (i.e., *false-belief* trials) or true (i.e., *true-belief* trials; Fig. 1c and Extended Data Fig. 2). All trial conditions were well-matched for difficulty and demand based on behavioral metrics for both other-belief vs. physical trials ($t = 1.04$, $p = 0.31$) and true-belief vs. false-belief trials ($t = 0.05$, $p = 0.96$; Fig. 1d and Extended Data Table 1b).

Many neurons in the dmPFC responded selectively when considering another's beliefs. Using linear models that quantified the degree to which the different conditions could be decoded from neuronal activity during questioning[23], we find that 20.0% ($n = 42$) of the neurons accurately predicted whether the participant was considering another's beliefs (other-belief vs. physical trials; permutation test, $p < 0.025$; Fig. 1e,f). Collectively, decoding accuracy for these neurons was 83 ± 2% and significantly above chance (Fig. 1g; permutation test, $p < 0.005$), suggesting that these neurons distinguished belief-related representations from other more generalized non-social representations.

In order to accurately infer the other's beliefs, it was necessary not only to consider another's beliefs but to also determine whether they were true of false. Here, we find that 23% ($n = 49$) of the neurons accurately predicted whether the participants were considering another's false vs. true beliefs (permutation test; p < 0.025; Fig. 2a, b). Collectively, the decoding accuracy for these neurons was 78±3% and significantly higher than chance (Fig. 2c; permutation test, p < 0.005). Similar findings were also observed when using other analytic techniques (Extended Data Fig. 3a, b), neural isolation approaches (Extended Data Fig. 3c), and time alignments (Methods) as well as when comparing decoding performances across the individual participants and clinical conditions (Extended Data Fig. 4). Neuronal responses to the other's beliefs were also robust to differences in cognitive demand (Extended Data Fig. 5a–c), complexity (i.e., number of social agents or relevant items; Extended Data Fig. 5d–f), and depth of reasoning required (i.e., first- vs. second-order beliefs; Extended Data Table 2b and Fig. 2d)[24]. Therefore, even though the social agents and context broadly varied across trials, these neurons appeared to reliably predict the other's beliefs.

Self vs. others' beliefs and perspective

It could be argued that neurons that were predictive of the other's beliefs may have simply signaled the presence of any inconsistency between past and present reality, irrespective of whether another's belief was involved. To test for this possibility, we required the participants to consider physical representations that were previously true but either currently false (i.e., *false-physical* trials) or true (i.e., *true-physical* trials; Extended Data Fig. 2). Here, we find however that, of the 49 neurons that distinguished between false vs. true beliefs, only 11 distinguished between false- vs. true-physical representations (permutation test, $p = 0.36$; Fig. 3a *left*). Moreover, by using neurons that were predictive of the other's beliefs and by employing models trained on true- vs. false-belief trials to decode true- vs. false-physical trials ('model-switching'; Methods), we find that decoding performances were at chance (50 ± 2% vs. 50% chance; permutation test, $p > 0.5$ Fig. 3a *right*), suggesting that these neurons reflected the other's beliefs of reality rather than its physical state.

To further confirm that neuronal predictions of the other's beliefs reflected the other's perspective of reality distinctly from the participant's own, we introduced an additional set of controls. On 20% of other-belief trials, we included *false-belief [aware]* trials in which the social agents were made explicitly aware of the critical manipulation events. For example, the participant may hear "…after Tom leaves, you move the jar to the cupboard as he watches you through the window." Therefore, when compared to the standard false-belief [unaware] trials, the other's beliefs were now true from the social agent's perspective (i.e., since they watched through the window). Here, we find that neurons that accurately predicted the other's beliefs on the standard other-belief [unaware] trials also accurately predicted the other's beliefs on other-belief [aware] trials (permutation test, $p < 0.001$; Fig. 3b). Moreover, considered collectively, decoding accuracies on these trials (false- vs. [aware] true-belief) were similar to those decoded from the standard other-belief trials (77 ± 2% vs. 78 ± 3% respectively; Permutation test, $p = 0.61$); Extended Data Fig. 6a) and positively correlated on a cell-by-cell basis (Pearson's correlation; $r = 0.3$, $p = 0.04$; Extended Data Fig. 6b). Neuronal predictions of the other's beliefs, therefore, reflected the other agent's specific perspective.

Finally, given these findings, we asked whether neurons that were predictive of the other's beliefs distinguished self- from other-belief related representations. While it is not possible for one to simultaneously hold a false belief and to know that one's belief is false, it is possible to evaluate how neurons may represent one's own imagined false beliefs. To test for such representations, we recorded from an additional 112 neurons while the participants judged their own beliefs as false or true (Extended Data Table 2b). Using these *self-belief* trials, we find 31 (27.7%) neurons that accurately predicted whether the participant's own imagined beliefs were true or false (Fig. 3c). These neurons, however, were also largely distinct from those that reflected the other's beliefs (permutation test, $p = 0.14$). Moreover, when using neurons that were predictive of the other's beliefs to decode the participant's own false vs. true beliefs (Methods), prediction accuracy on these self-belief trials was at chance (49 ± 3% vs. 50% chance; permutation test; $p = 0.33$; Fig. 3d). These findings therefore together suggested that neurons that were predictive of the other's beliefs were largely uninformative of the participant's own belief-related representations (i.e., they distinguished self- from other-related beliefs and perspective).

Predicting the other's beliefs contents

In order to provide the correct answer, it was necessary for the participants to infer not only whether the other's beliefs may be false but also the specific beliefs being considered. For example, whereas certain trials required the participant to consider another's beliefs of 'what' an item may be, other trials required them to consider 'where' it is located. Here, we find that 60% of the neurons accurately predicted whether the other's beliefs may be false about an item's identity and 34% accurately predicted their beliefs about its location (Fig. 4a and Extended Data Fig. 7). Moreover, when considered collectively, decoding accuracies for these neurons were similarly high at 81 ± 4% and 84 ± 4% respectively (Permutation test, $p = 0.15$), suggesting that these populations reliably encoded information about beliefs of what or where the items may be.

In order to accurately infer the other's beliefs, it was also necessary for the participants to determine the specific item being considered; for example, whether Tom believes the jar to be on the "table" or in the "cupboard". Therefore, to examine whether and to what degree the neuronal population may reflect such information, we further divided the items into six groupings – objects (e.g., table), containers (e.g., cupboard), foods (e.g., vegetables), places (e.g., park), animals (e.g., cat), and item appearances (e.g., red; Methods). Here, we find that the neural population predicted the specific item grouping with an accuracy of 64 ± 3% when compared to all other groupings (H_0 = 50% chance; permutation test, $p < 0.005$; Fig 4b). Moreover, when taken together, the probability of correctly predicting (i) whether another's belief was involved, (ii) whether the other's belief was false, (iii) whether the other's belief was of an item location or identity, and (iv) the specific item being considered was 36±2% (H_0 = 6.25% chance, permutation test; $p < 0.0001$; Fig. 4c). Therefore, when taken at the level of the cell population, these neurons appeared to encode highly detailed information about the content of the other's beliefs on a trial-by-trial level.

Population predictions and performance

Finally, we asked whether and to what degree the activities of these neurons reflected the participants' behavioral performances. Using a matched number of correctly and incorrectly answered trials, we find that population prediction accuracy of the other's false vs. true beliefs was 72% for correctly answered trials but only 56% for incorrectly answered trials (permutation test, $p < 0.005$; Fig 4d; see Extended Data Fig. 8 for other-belief vs. physical trials). A similar drop in decoding accuracy was also observed when considering all task-relevant features (37% vs. 13%; permutation test, $p < 0.001$), suggesting that the activities of these neurons correlated with the ability of the participants to correctly infer the other's beliefs. By comparison, we found no net difference in mean activity between correctly vs. incorrectly answered trials across any of the false-belief, true-belief, false-physical, or true-physical trials (one-way ANOVA, F(3, 2268) = 0.56, $p = 0.64$), suggesting that diminished decoding was not explained by more generalized processes such as lapses in attention or judgment. The participants' ability to accurately predict the other's beliefs was therefore reflected by the specific cell pattern of population activity on a *per*-trial level.

Discussion

By recording the activities of dmPFC neurons in participants performing a structured false belief task across richly varying naturalistic conditions, we observed that these neurons provided progressively granular levels of details about others' beliefs – from whether or not another's belief was involved, to whether these beliefs were true or false, to which particular item was being considered. Importantly, the activity of these cells distinguished another's beliefs from other non-social physical representations and disambiguated self- from other-belief representations and perspective; computations that together are essential to human theory of mind[3, 10, 25, 26].

These findings are notable because they reveal neurons in the human dmPFC that encode information about another's beliefs, even when those beliefs are false or distinct from one's own. Whereas canonical 'mirror neurons' in premotor and supplementary motor areas have been previously shown to reflect information about the observable behavior of others and to represent another's actions similarly to one's own [27, 28], it has remained unclear whether or how neurons represent another’s beliefs, which are inherently unobservable and unknown. Further, while neurons in non-human primates have been found to predict another’s actions or expected reward[29, 30], understanding whether or how individual neurons reflect another’s beliefs or perspective has largely remained out of reach. Here, we identify putative neurons in the human dmPFC that may support these computations.

A final notable finding from these studies is that single-cellular representations of the other’s beliefs were largely insensitive to differences in task difficulty or demand; reliably predicting the other’s beliefs across broadly varying social contexts and themes. They were also robust to differences in depth of reasoning required, suggesting that these neuronal representations of the other’s beliefs are likely generalizable; a property that would be necessary for supporting theory of mind. It was also notable to find, though, that many of the neurons encoded non-social information about the physical state of reality which could potentially explain why certain lesions in the dmPFC can lead to overlapping deficits, some of which are not necessarily specific to social reasoning[26]. Taken together, our observations provide a rare look into the cellular-level processing that underlie human theory of mind and a new understanding of how neurons in the human brain may reflect another’s beliefs, with prospective implication to human social cognition and its dysfunction[6–8].

Methods

Participants

Participant recruitment: All study procedures were performed under ethical standards provided by the Massachusetts General Hospital Internal Review Board and in compliance with Harvard Medical School ethical guidelines. Prior to consideration, candidates for the study were evaluated by a multidisciplinary team of neurologists, neurosurgeons, and neuropsychologists [31–34] and decisions for surgery were unrelated to study participation. Prospective candidates who displayed cognitive scores that lay outside 1.5 standard deviations of their age-defined means (e.g., WAIS-IV, WCST, and WMS-IV) were excluded [35, 36]. Consideration for inclusion in the study was only made after patients were scheduled for elective placement of deep brain stimulation. Their cases were reviewed for study candidacy based on the following inclusion criteria: 18 years or older, able to give informed consent, intact preoperative baseline language function and English fluency, and plan for awake surgery with intraoperative microelectrode recordings. All participants gave written informed consent to take part in the study. The patients were freely able to withdraw from the study without any consequence to their clinical care at any point in the study, including during the intraoperative phase.

A total of 15 participants were included for neuronal recordings (Extended Data Table 1). Of these, 11 participants (5 female and 6 male, mean age: 62 years, range: 32–73 years) underwent single-neuronal recordings while performing the main false-belief task. Of those participants, 7 had essential tremor (ET), 3 had Parkinson’s disease (PD) and 1 had dystonia. An additional 4 participants (3 female and 1 male, mean age: 54 years, range: 19–72 years) also underwent single-neuronal recordings while performing a false-belief task but that further tested for first-order vs. second-order false beliefs as well as self- vs. other-beliefs. Finally, a separate set of 14 healthy participants (age range 25–62 with 6 males and 8 females) were used to confirm the dependency between verbal response to the questions and the preceding narratives.

Neuronal recordings

Intraoperative single-neuronal recordings: Individuals undergoing deep brain stimulator placement at our institution normally undergo standardized micro-electrode recording as part of their clinically planned surgery in order to optimize anatomical targeting[32, 37]. Here, we adopted a surgical approach that allowed us to obtain acute single-unit recordings of this area as the micro-electrodes were advanced to target [31–33]. These recordings did not perturb the planned operative approach or alter clinical care (Extended Data Fig. 1a).

Neuronal recordings from the dmPFC was conducted in three main steps. First, to mitigate pulsations or movement at the cortical surface, we used a biodegradable fibrin sealant (Tisseel, Baxter; Deerfield, IL, USA), between the cortical surface and the inner table of the skull [2]. The sealant is normally used after deep brain stimulation placement but, in our setting, placement before micro-electrode targeting allowed for cortical pulsations to be additionally locally mitigated (Fig. 1a). Second, using a motorized microdrive, we incrementally advanced the micro-electrodes along the cortical ribbon at 10–100 μm increments in order to identify and isolate individual units (Alpha Omega Engineering, Nazareth, Israel). Here, we employed the same array of 5 tungsten microelectrodes (500–1000 kΩ) normally used for deep targeting. Once putative neurons were identified, the microelectrodes were held in position for 4–5 minutes in order to confirm signal stability (we did not screen putative neurons for task responsiveness). The electrodes were then left untouched until the end of the task session. The electrodes were then again advanced (by an additional 0.4 to 1.2 mm on average) until a different set of stable unit waveforms were obtained and another session commenced. Finally, a multielectrode recording (MER) system and I/O DAQ were used to precisely time-stamp task events (1kHz) and sample the neuronal data (44 kHz) at millisecond resolution. Neuronal signals were amplified, bandpass filtered (300 Hz and 6 kHz) and stored off-line (Alpha Omega Engineering, Nazareth, Israel). Audio recordings were obtained at 22 kHz sampling frequency using two microphones (Shure; Niles, IL USA) that were integrated into the Alpha Omega rig for high fidelity temporal alignment with neuronal data. After recordings from the dmPFC, subcortical neuronal recordings and deep brain stimulator placement proceeded as scheduled.

Single-unit isolation: Single-units were identified and sorted off-line (Plexon offline sorter, Plexon Inc. Dallas, TX). To ensure the identification of single, well-isolated units, we first constructed a histogram of peak heights from the raw voltage tracings on each channel. A minimum threshold of three standard deviations was used to differentiate between neural signals from background noise. Next, template matching and principal component analyses were used to classify action potentials and sort prospective neurons. Candidate clusters of putative neurons needed to clearly separate from channel noise (>3 s.d. above baseline), display a voltage waveform consistent with that of a cortical neuron, and to have at least 99% of action potentials separated by an inter-spike interval of at least 2 msec. Any prospective units that displayed significant overlap in their principal component analysis

(PCA) distributions by multivariate analysis of variance ($p < 0.0001$) or overlapped with the baseline signal/noise were excluded from the single-neuronal analysis. Finally, any units that did not demonstrate waveform stability over the course of the trial were excluded from further analysis. Extended Data Fig. 1b, c illustrates two examples of spike waveform morphologies and associated PCA clusters. In total, we recorded from 212 putative neurons across 17 recording sessions for an average of 1.5 recording sessions *per* participant in the main task. The average number of neurons isolated per recording session was 12 ± 1, amounting to approximately 2 well-isolated units *per* electrode *per* session across the 5 recording electrodes [33, 34, 38, 39].

Multi-unit isolation: To provide further comparison to our single-neuronal data, we also separately analyzed multi-unit activity (MUA). MUAs represent the combined activities of multiple neurons from within local populations recorded from the same electrode. Here, as described previously [40, 41], MUAs were isolated from the same electrodes in which single-units were isolated. Like single-unit activity, they were separated from noise by baseline thresholding but, unlike single-unit activity, they were not processed for waveform morphology or separability.

Audio processing: Audio recordings were obtained at 22 kHz sampling frequency and time-aligned to spiking activity using the Alpha Omega recording system. The starting and ending time points of each narrative and question were then annotated using WaveSurfer software (KTH Royal Institute of Technology, Sweden). Each word that was heard (i.e., narrative and question) and spoken (i.e., answer) was then manually transcribed and confirmed for alignment using custom written software in MATLAB (MathWorks, Inc.; Natick, MA, USA). Finally, the narratives, questions, and answers were tabulated based on whether and what type of belief was being considered (e.g., self-belief trial, other-belief trial, etc.; Extended Data Table 2).

Task design

The behavioral task was administered in an automated fashion using customized software written in MATLAB. After stability of neuronal recording was confirmed, the patients were then given, in auditory format, varied narratives followed by questions about them over multiple trials. To ensure that the presentation of the narratives was naturally ‘blinded’, the narratives and questions were pre-recorded in audio and were given to the participants *via* computer. The narratives lasted, on average, 7.68 ± 0.07 seconds *per* trial and described simple events such as an object being moved from one location to another or a box being opened, and the questions focus on the state of the objects or a social agent’s belief of them. Therefore, in one trial, the participant may be given a narrative such as “You and Tom see a jar on the table. After Tom leaves, you move the jar to the cupboard”. This would then be followed by the question “Where does Tom believe the jar is?” Other narratives, by comparison, may present a scenario such as “You placed an apple inside a shoebox while Sallie was not watching. Sallie then opens the shoebox” followed by the question “What does Sallie expect to find in the box”. To further allow for generalizability, we also alternated between words such as “Tom” and “he” or “think” and “believe” during questioning. Overall, the participants were given 144 ± 22 unique narrative and question combinations, with the average question duration being 1.96 ± 0.15 seconds *per* trial

To evaluate for neuronal responses that may selectively reflect another's beliefs, it was important to dissociate information relevant to the scenarios within the narratives from the information being considered during the questioning period. It was also important to prevent the participants from using simple learning strategies when given particular story scenarios to anticipate which question will be given. To this end, we also randomized the type of questions following the narratives. For example, certain trials would present the narrative "You and Tom see a jar on the table. After Tom leaves, you move the jar to the cupboard". Whereas some trials would be followed by the question "Where does Tom believe the jar is?", other trials would be followed by the question "Where do you think the jar is?" Specific narrative-question variations and controls are given further below and in Extended Data Table 2a.

Primary task conditions

In our study, we used the story narratives and questions about them to vary the content and theme under which the participants had to consider another's belief (Extended Data Fig. 2). As detailed below, we also used them to test for changes in neuronal activity that may reflect information specifically related to another's beliefs as well as to evaluate for features that describe another's belief in progressively granular details.

Other-belief vs. physical trials: To evaluate for neurons that responded selectively when considering another's beliefs, we compared trials that required the participant to consider another's beliefs of reality vs. those that required them to consider its physical representation. For example, the participant may be presented with the narrative "You and Tom see a jar on the table. After Tom leaves, you move the jar to the cupboard" followed by the question "Where does Tom believe the jar is?" These other-belief trials would therefore require the participant to consider the other's belief during questioning. Other trials, by comparison, would present the participant with the narrative "You take a picture of a jar on the table. After the picture, you move the jar to the cupboard" followed by the question "Where is the jar in the picture?" These physical trials would therefore require the participant to consider the physical state of reality and would not involve another's beliefs.

True- vs. false-belief trials: Next, to identify putative signals that may be predictive of others' specific beliefs, other belief trials were further divided into those that required the participant to consider beliefs that were false vs. those that required them to consider beliefs that were true. Thus, for example, on false-belief trials, the participant may be presented with the narrative "You and Tom see a jar on the table. After Tom leaves, you move the jar to the cupboard" followed by the question "Where does Tom believe the jar is?" On true-belief trials, by comparison, they would be presented by the narrative "You and Tom see a jar on the table. After Tom leaves, you open the jar and leave it in place." This would then be followed by the question "Where does Tom think the jar is?" Therefore, even though the questions for both trials are the same, only the former reflects a belief that is false.

True- vs. false-physical trials: To test for the possibility that neurons encoding others' beliefs may have simply signaled the presence of any inconsistency between past and present reality, irrespective of whether another's belief was involved, similar orthogonal manipulations were made for physical trials. Whereas certain trials involved other's beliefs that were true vs. false other trials involved physical representations that were true vs. false. For example, certain trials may contain a narrative such as "You take a picture of a jar on the table. After the picture, you then move the jar to the cupboard". Other trials, by comparison, may contain a narrative such as "You take a picture of a jar on the table. After the picture, you open the jar and leave it in place". Therefore, whereas the former trial requires the participant to consider a false physical-representation of reality when asked "Where is the jar in the picture?", the latter trial requires them to consider a true physical-representation.

Additional task variations and controls

Other-belief aware vs. unaware trials: On false-belief trials, the social agents held representations of reality that were false and distinct from the participant's own because the agents were unaware of events. For example, when presented with the scenario "...After Tom leaves, you move the jar to the cupboard", Tom is not aware that the jar was moved. Therefore, to evaluate whether neuronal responses reflected variations in the other's perspective of reality independently of the participant's own, we introduced an additional set of control trials in which the social agent's awareness was implicitly varied (20% of other-belief trials). For instance, whereas certain trials contained narratives such as "...After Tom leaves, you move the jar to the cupboard," other trials contained narratives such as "...After Tom leaves, you move the jar to the cupboard while he watches through the window." Therefore, even though both trials describe the same manipulation events (e.g., moving the jar to the cupboard), the social agent in the latter is implicitly aware of them.

Self-belief vs. other-belief trials: While it is not possible for one to simultaneously hold a false belief and to know that one's belief is false, it is possible to evaluate how neurons may respond to representations of one's own imagined false beliefs. Therefore, to evaluate for neurons that may distinguish self- from other-belief related representations, the participants were given trials in which their own belief had to be judged as false or true. For example, the participant may be given the narrative "You see a jar on a table. After you leave the kitchen, the jar falls off the table onto the floor." followed by the question "Where will you expect to find the jar?" (Extended Data Table 2b).

First-order vs. second-order other belief trials: While first-order false beliefs require the participant to consider another's beliefs, second-order false beliefs require the participant to consider another's beliefs of another's beliefs [42]. For example, on a second-order false belief trial, the participant may be presented with the narrative "Mary and Tom see a jar on a table. Tom leaves the kitchen and Mary moves the jar to the cupboard. Tom returns." followed by the question "Where does Mary think Tom will look for the jar?" (Extended Data Table 2b). Therefore, while both require consideration of a false belief, the latter involves a higher depth of reasoning and task demand. Here, we theorized that, if our results were explained by a difference in depth of reasoning or difficulty, then we should expect to find differences in neuronal activity or decoding accuracy when comparing these first- and second-order belief trials.

Other-belief trials for item identity vs. location: To examine the consistency of neuronal response across belief contents, the participants had to consider another's beliefs about either an item's location or its identity; thus, trials were divided into two groups accordingly. Therefore, for trials that required the participant to consider an item's location, they may be given a narrative such as "You and Tom see a jar on the table. After Tom leaves, you move the jar to the cupboard" followed by the question "Where does Tom believe the jar is?" Other trials, by comparison required the participant to consider the item's identity. Here, for example, they may be given a narrative such as "You and Tom see a jar on the table. After Tom leaves, you replace the jar with an apple" followed by the question "What does Tom believe is on the table?"

Other-belief trials for item groupings: To investigate whether neuronal signals may reflect the specific content of the other's beliefs, we varied the items being considered by the social agents in the narratives. When asked "Where does Tom think the jar is?", for example, the participant had to correctly infer that Tom believed the jar to be on the "table" rather than "cupboard". Therefore, in order to further evaluate whether and to what degree neurons in the population may be informative of the items being considered, we divided the items into six groupings. These included common objects (e.g., chair), containers (e.g., cupboard), food items (e.g., vegetables), places (e.g., street), animals (e.g., cat) and appearances (e.g., red). For example, when asked "What does Jim believe is in the garden?" the participant had to consider "vegetables" which are a food item whereas, but when asked the question "Where will Ned look for the car?" they had to consider "street" which is a place.

Confirming the dependency between questions and narratives

The questions given to the participants allowed us to probe for specific information about social agents described in the narratives and their beliefs. Therefore, to confirm that the participants could not guess the correct answers from the questions themselves, we also presented questions without the preceding narratives in a separate set of controls. Here, we presented subjects with the same precise pre-recorded questions used for the main task. These were then followed by two forced-choice options of what the possible answers could be. Thus, for example, they may hear the question "Where does Tom think the jar is?" followed by the two options "table" or "cupboard". Using 14 healthy controls (age range 25–62 with 6 males and 8 females), we find that the participants selected the correct answer on only 52.4 ± 2.0% of the questions. Given a chance probability of 50%, the likelihood of answering correctly without hearing the narratives was therefore at chance (*t*-test, $t(13) = 1.2$, $p = 0.25$).

Statistical Analysis

Single-neuronal analysis: Neuronal activity was analyzed during the question period, at which time the participants were considering the specific information being asked. To standardize neuronal analysis and to take into account the known time delay between stimulus presentation and neuronal response by prefrontal neurons [31, 34], we focused on a

1000 msec window starting 200 msec from question onset. To construct the peri-stimulus time histograms (PSTH), the spike train of each unit was first converted to a continuous spike-density function (SDF) using a Gaussian smoothing kernel with width of 100 msec [43, 44]. To allow for consistency across trials, the firing activities were aligned to the question onset.

A Fisher discriminant was used to evaluate whether and to what degree the activity of each neuron during questioning could be used to predict specific trial conditions on a trial-by-trial basis [45–48]. A permutation test was used to evaluate for statistical significance (permutation test, $p < 0.025$) and Bonferroni corrected for other-belief (false vs. true) and physical representation (false vs. true) comparisons. As described previously[10], the ratio of the variance in neuronal activity between the two groups of trials was compared to the variance within groups based on:

$$S_W^{-1} S_B v = \lambda v,$$

whereby S_W and S_B are the within group scatter matrices and between group scatter matrices, respectively. The prediction vector v, corresponds to the largest eigenvalue of the matrix on the left-hand side of the equation. The prediction vector defines a projection of the recorded activity into a scalar unit that is then compared to a threshold, θ; for example, the trial type was predicted to be ‘false-belief’ if greater than θ and ‘true-belief’ if less than θ. For validation, we divided the neuronal data into a training set consisting of 80% of the trials and tested the accuracy of the prediction on the remaining 20% of trials. This operation was repeated 200 times using a random sampling of the total trials. A chance distribution of decoding performance was also generated using the same procedure while randomly shuffling the labels corresponding to each trial (e.g., randomly shuffling true-belief with false-belief trials). A decoding performance of 100% therefore indicates a perfect prediction whereas a decoding performance of 50% indicates chance. Finally, to visualize the temporal structure of the decoding accuracy over the course of the trials, we performed a sliding window analysis. Here, we used a sliding window of 1000 msec moving in steps of 100 msec from −1500 to 2500 msec relative to the question onset.

Model-switch decoding: To quantify the degree to which neuronal responses are selective, we used a model-switch procedure whereby models trained on certain trial conditions were used to decode a different trial condition on validation trials not used for model training. For example, to test for the selectivity of the neuronal response to another’s beliefs, we would train models on false- vs. true-belief trials and then use these models to decode false- vs. true-physical trials. Therefore, even though both trial conditions involve false vs. true representations, a drop in decoding accuracy on model-switching would suggest that neuronal responses were selective for another’s beliefs.

Neural population decoding analysis: To further evaluate whether and to what degree the activity of the neuronal population was informative of the trials being given, we again used Fisher discriminant but now constructed a pseudo-population activity matrix ($m \times n$) of neurons of interest. Each cell in the population activity matrix contained the mean firing rate from a single neuron *n* on a single trial *m* measured during the task. Only neurons with a minimum number of 10 trials per trial type were included in the analysis. Because neurons were not simultaneously recorded, trials from different neurons were randomly matched up according to their trial type (e.g., false-belief or true-belief trials). This procedure was repeated 200 times with different random trial matching. Similar to the procedure used for the individual cells, the data was split into a training set consisting of 80% of trials and tested on the remaining 20% of trials for validation. We also balanced the number of trials from each condition for training and testing. Population decoding accuracy was then quantified as the percentage of correctly classified trials, averaged across all 200 iterations of random trial matchings. A chance distribution of decoding performance was also generated using the same procedure while randomly shuffling the labels corresponding to each trial. As before, the decoding performance of the neuronal population was considered significant if its average performance fell within the top 2.5% of the chance performance (p-value < 0.025). Lastly, to investigate the contribution of the cumulative population, we randomly selected k neurons ($k = 1, 2, 3, \ldots n$ where n is the overall population size) at each step and then determined the average decoding performance by repeating this procedure 200 times. Moreover, since our approach ignores the potential contributions from cross-correlations between neurons, it likely provides a lower bound for decoding performance.

Statistical validation: A parametric *t*-test and a non-parametric rank-sum test were further used to validate the significance and magnitude of the neuronal response. Here, rather than evaluating the probability of correctly decoding the trial conditions compared to chance, we evaluated the statistical significance of neuronal response across conditions ($p < 0.025$). To further evaluate the magnitude of the effect, we also calculate the *t*-statistic and *z*-value metrics over the course of the trials. Here, similar to our decoding approach, we used 1000 msec sliding windows that were incrementally advanced in steps of 100 msec but now calculated the *t*-statistic and *z*-values (Extended Data Fig. 3a, b).

Trial complexity analysis: To evaluate the potential relation between neuronal activity and trial difficulty and demand, we used three standard complexity measures: (i) the number of relevant items considered during the narratives, (ii) the number of times an agent was considered in the narratives, and (iii) the narrative length. For the number of relevant items, we considered the number of items that had to be held in the working memory prior to the questioning (e.g., 3 for jar + table + cupboard vs. 4 for street + bicycle + car + garage in Extended Data Table 2). For the number of social agents, we counted any instantiation of an individual. Overall, the number of times that a social agent was mentioned within other-belief and physical-representation trials was well-matched (3.3 ± 0.1 vs. 3.4 ± 0.1 agents, respectively; rank-sum test: z-value = 0.68, $p = 0.49$).

Trial difficulty analysis: To investigate the perceived difficulty of the questions across the participants, we divided the trials into those that were considered easy vs. hard based on the participants' performances. We also divided the trials into those in which the reaction times between question offset and answer onset was short vs. long and found no relation between neuronal activity and the reaction times of the participants (i.e., short vs. long; rank-sum test:

z-value = 1.12, $p = 0.26$). The divisions were defined based on the median values across all trials and participants.

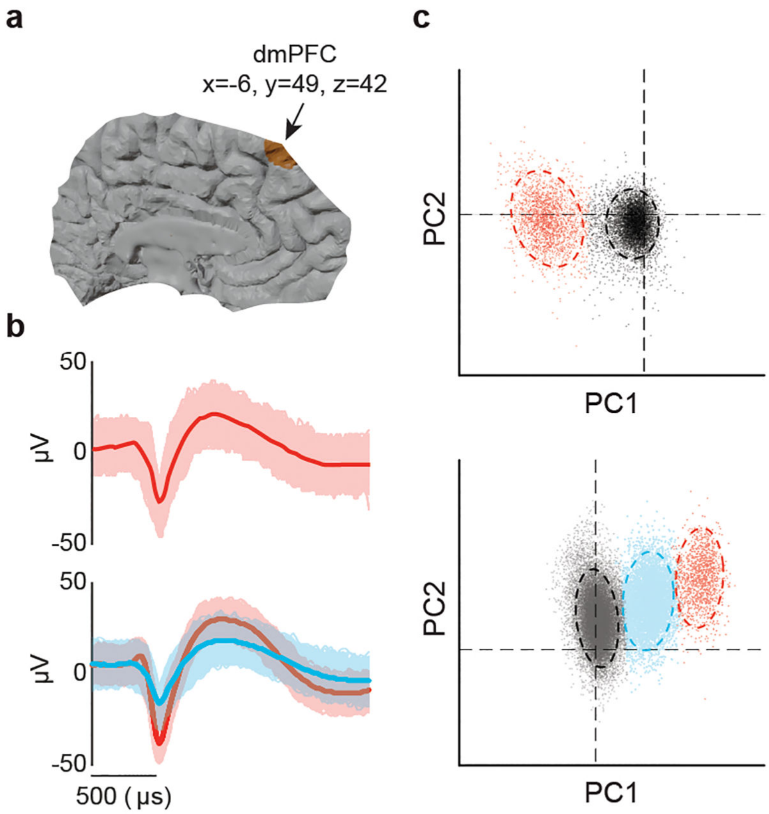
Trial uncertainty analysis: A minority of trials used in our main false-belief task ($n = 6$ out of a total of 95 narratives; Extended Data Table 2b) required some degree of inferences about the location of the item of interest when answering the questions that were not explicitly stated within the narratives. For example, when the participants hear, "Ned and you left a car in the street and a bicycle in the garage. While Ned was sleeping, you switched them. Tomorrow" followed by the question "Where will Ned look for the car?" The location of the car is not explicitly mentioned in the narrative and requires inference through the meaning of the verb "switch". Importantly, we found that the neuronal decoding was robust to differences in the degree of inference required [33, 49]. To this end, we repeated the decoding analysis after excluding the high-inference trials and found no difference in decoding accuracy for false vs. true beliefs based on whether the trials involved more or less uncertainty (78±3% vs. 77±2% prediction accuracy; $t = 0.76$, $p = 0.45$).

Consistency of neuronal encoding during questioning: To examine how question time-progression influenced neuronal encoding, we aligned neuronal activity to different time points during questioning. First, we aligned neuronal activity to the specific word at which sufficient information was given to correctly answer the question. Thus, for example, when hearing the question "Where does Tom think the pencil is?", the word "pencil" would be tagged as the word of interest. These words were selected through a natural language processing module that identifies their dependencies using a long short-term memory (LSTM) artificial recurrent neural network and parts-of-speech tagging [50, 51]. By aligning neuronal activity to the words of interest, we find that decoding for other-beliefs vs. physical representations was 72±2% and significantly above chance ($H_0 = 50\%$ chance probability, permutation test, $p < 0.005$). Similar findings were also made when evaluating decoding performances for false- vs. true-belief trials, with a decoding accuracy of 77±2% ($H_0 = 50\%$ chance probability, permutation test, $p < 0.005$). We also aligned neuronal activity to the end of the questions. Here, we found that prediction accuracy for the population was slightly lower at 68±2% for other-beliefs vs. physical representations and 74±2% for false vs. true beliefs ($H_0 = 50\%$ chance probability, permutation test, $p < 0.005$). Neuronal predictions about the other's beliefs therefore appeared to largely peak once sufficient information was available (on average) to comprehend and provide the appropriate answer.

Consistency of neuronal encoding across sessions: To rule out the possibility of habituation and to confirm the consistency of behavioral performance and neuronal decoding over time, we compared the first and second sessions. Most participants performed 2 sessions (1.5 sessions on average). Overall, we find no difference in the participant's performance when comparing the first to second sessions (81.2 ± 6.3% vs, 78.9 ± 11.0%; two-sided paired *t*-test, $t(4) = 0.45$, $p = 0.68$). We also find a similar proportion of task-modulated neurons when considering belief vs, physical representation trials (session #1: 23 ± 8% vs. session #2: 25 ± 6%; two-sided paired *t*-test, $t(4) = 0.70$, $p = 0.52$) as well as false vs. true belief (session #1: 27 ± 3% vs. session #2: 24 ± 5%; two-sided paired *t*-test, $t(4) =$

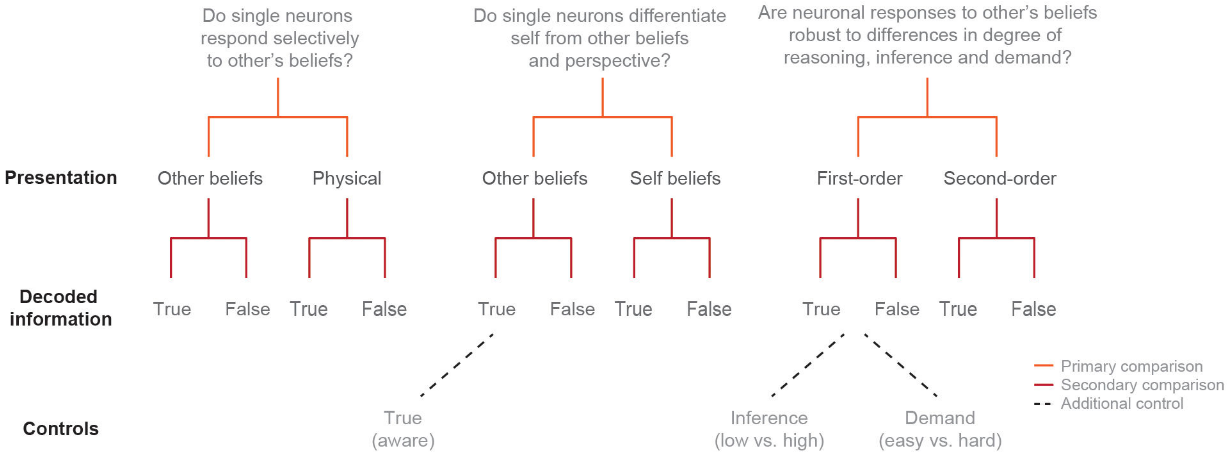
0.45, $p = 0.67$). Both behavioral performance and neuronal encoding were therefore consistent across sessions.

Extended Data



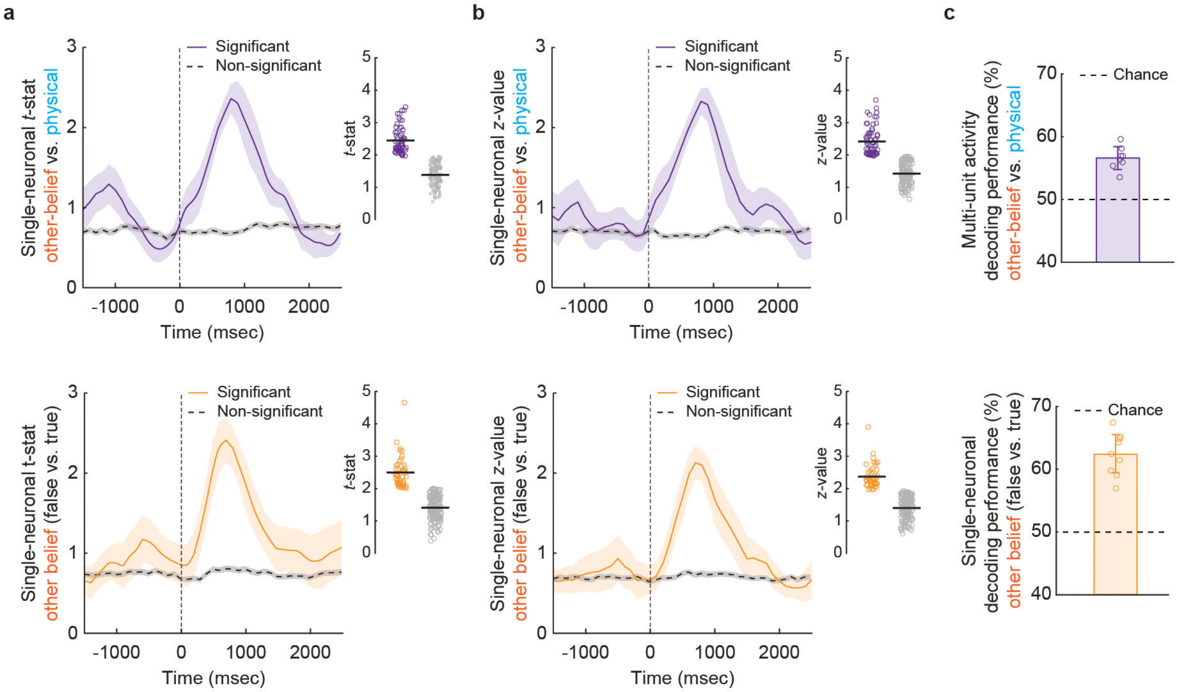
Extended Data Figure 1 |. Recording location, waveform morphology and single-unit isolation.

a, Single-neuronal recordings were obtained from the superior frontal gyrus of the dmPFC using incrementally advancing microelectrode arrays. The region of recordings in MNI coordinates ($x = -6$, $y = 49$, $z = 42$) is shown in a canonical structure MRI. **b,** Examples of waveform morphologies displaying mean waveform ± 3 standard deviations. The *top* panel illustrates a single representative unit isolated from a fine-tip tungsten microelectrode. The *bottom* panel illustrates two representative units that were isolated from another microelectrode. The horizontal bar indicates a 500 μs interval for scale. **c,** Isolation patterns corresponding to the waveforms shown in **c** represented by principal component distributions. The gray areas in the PC space represent baseline noise. All putative units displayed significant separation by one-way MANOVA ($p < 0.0001$) and no overlap with baseline signal/noise.



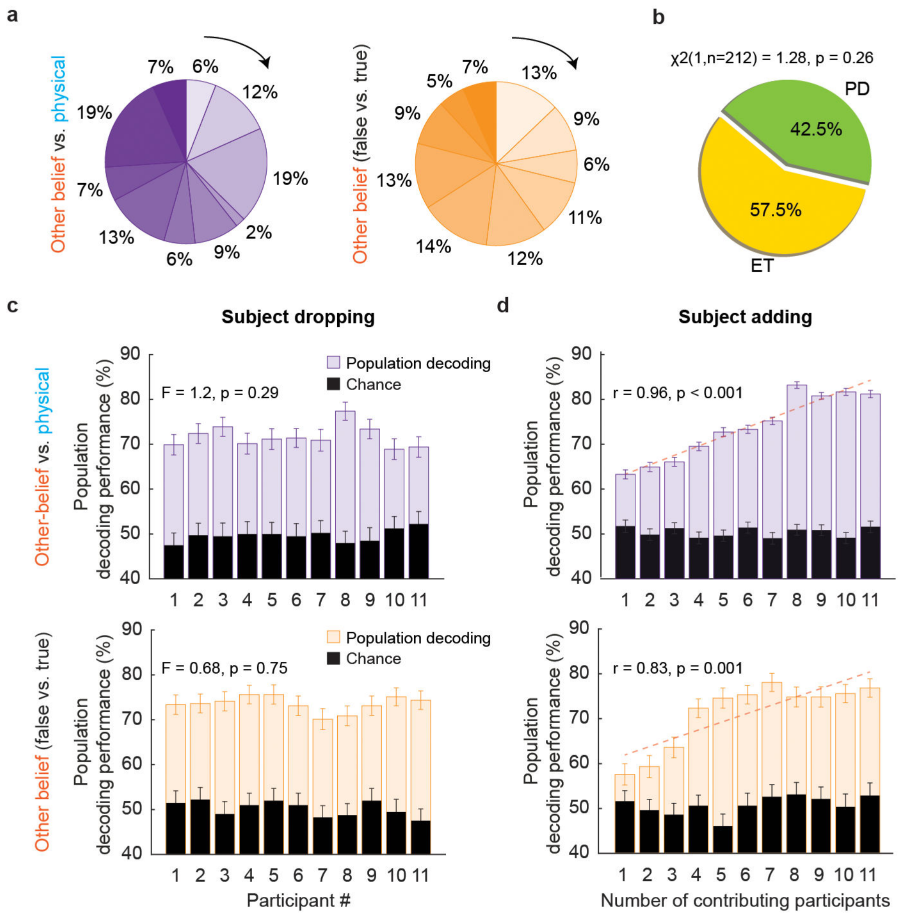
Extended Data Figure 2 |. Schematic depiction of experimental logic and narrative features across trial conditions.

On the *left*, other belief vs. physical trials were used to identify neurons that responded selectively to another's beliefs. Whereas both required the participant to consider false vs. true representations, only the former required the participants to consider another's specific beliefs. In the *middle*, other belief vs. self-belief trials were used to further differentiate other- from self-related representations. Whereas both required the participant to consider a belief, only the former required the participants to consider another's false vs. true beliefs. Aware vs. unaware trials were given to additionally differentiate other- from self-perspective. On the *right*, first- vs. second-order belief trials were used to evaluate for the consistency of neuronal response across different depths of reasoning. High vs. low degree of inference as well as high vs. low task demand trials were used to evaluate for the consistency of neuronal response across different degrees of inference and cognitive demand.



Extended Data Figure 3 |. Consistency of the results across different statistical methodology and neuronal isolation approaches.

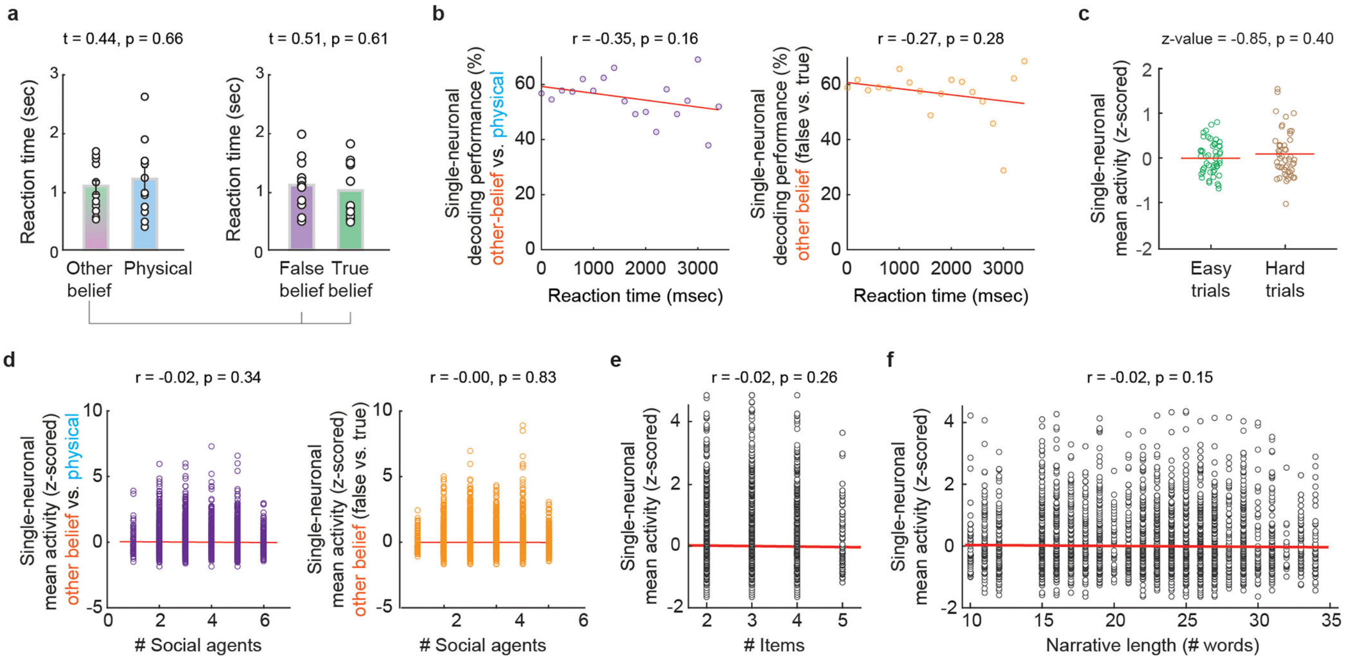
a, A parametric two-sided unpaired *t*-test was used to evaluate whether cells displayed a significant difference in their responses. Comparisons were made between other belief vs. physical trials (*top, n* = 62 neurons) and between false vs. true other-belief trials (*bottom, n* = 47 neurons). The magnitude of effect (mean ± s.e.m) over the course of the trial is displayed based on the *t*-statistic. Neuronal activity is aligned to the question onset (time zero). The *insets* display the *t*-statistic values for all neurons that displayed (*n* = 62 in the *top* and *n* = 47 in the *bottom* panel, colored) and did not display (*n* = 150 in the *top* and *n* = 165 in the *bottom* panel, gray) significant selectivity. **b,** A two-sided unpaired non-parametric rank-sum test was used with the same conventions as above. Here, the magnitude of effect (mean ± s.e.m) is displayed based on the *z*-value (*n* = 64 in the *top* and *n* = 45 in the *bottom* panel**)**. **c,** These results also held when considering other neural isolation approaches. Decoding performances were obtained for multi-unit activity (MUA) using the same modeling and decoding approach as for the single-neuronal data. The bar graphs provide the individual MUAs (*n* = 8) and their corresponding 95% CL. The horizontal line indicates chance performance (one-sided permutation test, $p < 0.005$).



Extended Data Figure 4 |. Consistency of the results across subjects and clinical conditions.

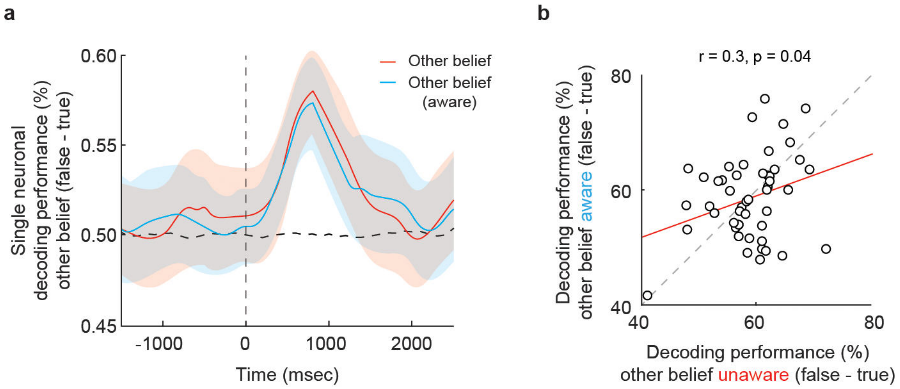
a, The participants demonstrated a largely similar proportion of task-modulated neurons when considering belief vs. physical trials (s.d., of 11.6%) as well as false- vs. true-belief trials (s.d., of 10.0%). **b,** Proportion of neurons displaying task modulations based on clinical conditions; Parkinson's disease (PD) and essential tremor (ET). The *p*-value by chi-square test is shown. We also found no difference in the firing rates of the neurons based on clinical diagnosis (1.61±0.19 vs. 1.70 ± 0.11 spike/sec for PD and ET, respectively; two-sided Wilcoxon rank-sum, *z*-value (1586) = 0.92, *p* = 0.36). **c,** A subject-dropping procedure was used to determine whether any of the participants disproportionately contributed to the population decoding performance. Here, individual participants were sequentially removed one at a time and the population decoding was repeated (200 iterations). Population decoding performances (mean ± s.e.m.) are separately presented after each participant was removed. Chance decoding based on random permutation of the neuronal data is provided in black for comparison. The decoding performances were largely unaffected by removal of any of the participants when decoding other-beliefs vs. physical representations (*top* panel; one-way ANOVA: *F*(10,2189) = 1.2, *p* = 0.29) as well as when decoding other true- vs. false-beliefs (*bottom* panel; one-way ANOVA: *F*(10,2189) = 0.68, *p* = 0.75). **d,** A subject-adding procedure was further used to determine how the participants cumulatively contributed to the population decoding by sequentially adding subjects contributing to the neuronal population from 1 to 11 and repeating the decoding analysis (200 iterations).

Decoding performances are provided with the same convention as above (mean ± s.e.m.). As shown, adding subjects one at a time led to a consistent increase in the decoding performance suggesting that the participants made similar contributions.



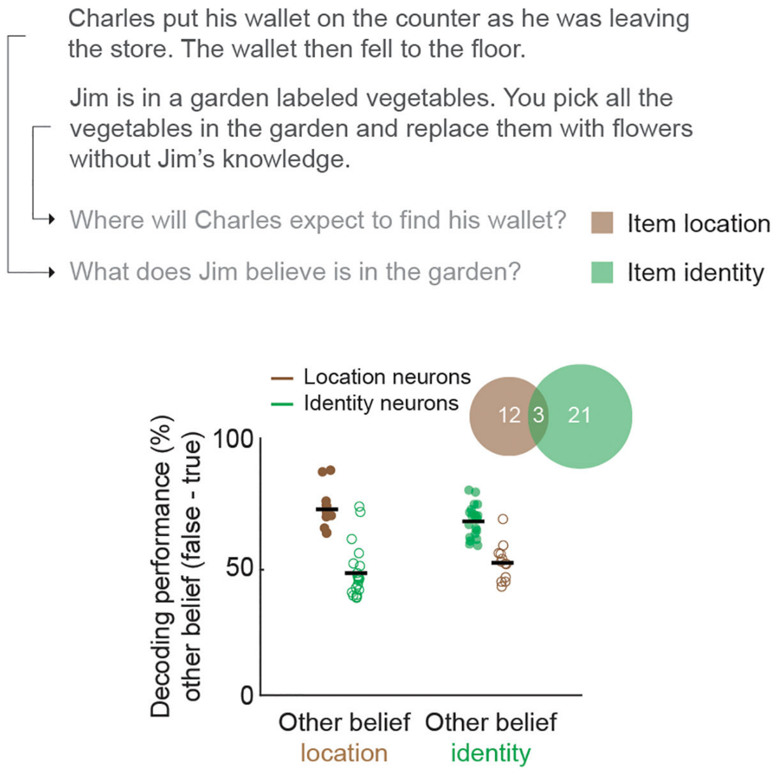
Extended Data Figure 5 |. Robustness of belief representations.

a, Reaction times (mean ± s.e.m.) from question offset to answer onset during the primary task conditions across participants (n = 11) were similar for other-belief vs. physical trials (1071 ± 135 vs. 1178 ± 201 msec) and for false- vs. true-belief trials (1130 ± 136 vs. 1028 ± 147 msec). The p-values obtained using a two-sided unpaired t-test. **b,** To evaluate how differences in neuronal decoding may relate to answer response time, decoding performances were first averaged across neurons that displayed significant selectivity and then sorted based on the participants' reaction times (n = 18 time points). There was a slightly negative but non-significant correlation between RTs and decoding performances both when comparing other-belief to physical trials ($r = -0.35$, $p = 0.16$) and when comparing other false- to true-belief trials ($r = -0.27$, $p = 0.28$). The p-values by Spearman's correlation test are shown. **c,** We found no relationship between neuronal activity (mean firing rates, n = 49 neurons) and trial difficulty (easy vs. hard; Methods) based on the participants' overall performances (two-sided rank-sum test, z-value = 0.85, p = 0.40). **d,e,f,** Neuronal activity was evaluated based on (**d**) the number of social agents presented to the participants (*left*: n = 4024 trials, *right*: n = 4527 trials), (**e**) the number of items (e.g., *table*, *jar*, *cupboard*, etc.) that had to be held in working memory prior to questioning (n = 4527 trials), and (**f**) the narrative length based on the number of words (n = 4527 trials). Activities were z-scored by removing the mean and dividing by the standard deviation. A lack of relation was demonstrated by correlation analysis in each condition (Pearson's correlation, $p > 0.1$).



Extended Data Figure 6 |. Decoding others' beliefs based on variations in perspective and awareness.

a, Mean decoding profile with 95% CL for all neurons that accurately differentiated between false-belief vs. true-belief trials ($n = 49$; one-sided permutation test, $p < 0.025$). Here, the trials were divided based on whether or not the social agent was made aware of events in the narratives. Since the state of reality was the same under these two conditions, demonstration of similar decoding performances on the standard other-belief and other-belief aware trials confirmed that neuronal predictions of the other's beliefs reflected the other's perspective of reality independently of the participant's own. **b,** Decoding accuracies on other-belief aware trials were positively correlated with those decoded from the standard other-belief trials on a cell-by-cell basis ($n = 49$; Pearson's correlation; $p = 0.04$).


Charles put his wallet on the counter as he was leaving the store. The wallet then fell to the floor.

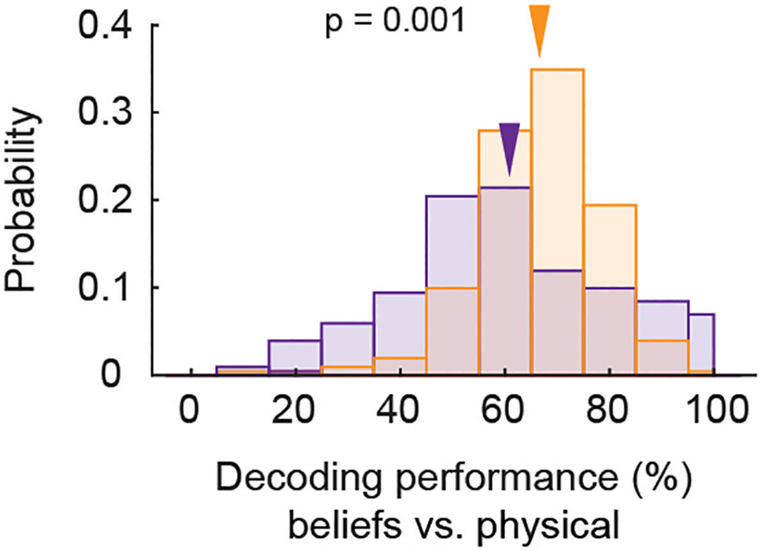
Jim is in a garden labeled vegetables. You pick all the vegetables in the garden and replace them with flowers without Jim's knowledge.

Where will Charles expect to find his wallet?

What does Jim believe is in the garden?

Extended Data Figure 7 |. Decoding others' beliefs based on variations in the item's identity or location being considered.

Above, the narratives and questions were varied in whether they required the participants to consider an items location or its identity. *Below,* the decoding performances of the individual neurons based on whether the social agent's beliefs involved an item's identity or its location are displayed. The Venn diagram (*inset*) shows the overlap between neurons.



Extended Data Figure 8 |. Relation between neuronal predictions and performance for beliefs vs. physical trials.

The histograms indicate decoding accuracies for neurons that predicted whether the participants were considering another's beliefs vs. physical representations on trials in which the participants provided the correct vs. incorrect upcoming answer (one-sided permutation test, $p = 0.001$). The arrows indicate mean decoding performances.

Extended Data Table 1 | Participants' demographics and overall performances.

a, Table summary displaying each participant's age, sex, underlying pathology, and number of trials per task per session. **b,** Overall, we find that participants with PD and those with ET displayed no difference in behavioral performances (mean ± s.e.m) across task conditions (two-sided unpaired *t*-test, $t(10) = 0.57$; $p = 0.59$). Analogous observations were also made when comparing the participant's mean reaction times from question offset to answer onset; with the participants with PD displaying largely similar response times to those of participants with ET (987 ± 56 vs. 1002 ± 41 msec; two-sided unpaired *t*-test, $t(1586) = 0.20$, $p = 0.84$). Finally, we found no correlation between participants' age and performance (younger vs. older than 65 years, two-sided unpaired *t*-test, $t = -0.07$, $p = 0.94$), together suggesting that clinical condition or age had no effect on the participant's ability to perform the task.

a

Case	Sex	Age	Diagnosis	# Sessions	# Trials per session

Primary tasks					Belief	Phy rep.
1	M	72	ET	1	47	50
2	F	73	ET	2	50	50
3	F	72	ET	1	49	49
4	F	32	ET	3	50	50
5	M	65	ET	2	38	38
6	M	51	PD	1	36	36
7	F	59	DYS	2	50	50
8	M	64	ET	2	50	49
9	F	68	ET	1	47	45
10	M	54	PD	1	45	44
11	M	72	PD	1	46	46
Additional control tasks					Self-belief	Other belief
1	F	54	PD	1	78	78
2	M	19	ET	2	59	59
3	F	72	PD	1	48	48
4	F	71	PD	1	52	52

b

	n	Performance (%)	*p*-value	# Neurons
Primary tasks				
Participants	11	81.0 ± 3.7	–	212
Age				
<= 65 yrs.	6	81.1 ± 4.5	0.94	145
> 65 yrs.	5	80.6 ± 6.2		67
Diagnosis				
Essential tremor	7	82.4 ± 5.3	0.59	156
Parkinson’s disease or dystonia	4	77.7 ± 4.7		56
Additional control tasks				
Participants	4	76.1 ± 3.8	–	112

ET: Essential tremor

PD: Parkinson’s disease

DYS: Dystonia

Extended Data Table 2 | Narrative and question examples.

a, Representative examples of narrative and question combinations that were given to the participants during neuronal recordings. To allow for generalizability, the narratives were varied in content and theme and the questions differed in the way they were asked (e.g., whereas certain trials asked “What will Tom...” other trials asked “What will he...”). Representative examples are given for both other-belief and physical trials. Additional trial variations and controls are described in the main text and Methods. **b,** Representative examples of narrative and question combinations that were given to the participants in order to additionally test for (i) self-belief related representations, (ii) second-order belief representations, and (iii) differences in the degree of inference required.

a	***Primary tasks***	
	Narrative examples	Question examples
	You and Tom see a jar on a table. After Tom leaves, you move the jar to the cupboard.	Where will he expect to find the jar?
	You and Jim take a photo of a bird on a tree. While the photo develops, the bird flies to a nearby rock.	Where is the bird in the photo?
	Ned left a car in the street and a bicycle in the garage. While Ned was sleeping, the car and bicycle were switched.	Where will Ned look for his car?
	You and Mary are at the movie theater sharing popcorn out of a container. She leaves and you eat some of the popcorn.	What does Mary believe is in the container?
	Charles put his wallet on the counter as he was leaving the store. The wallet then falls to the floor.	Where will Charles expect to find his wallet?
	John and you are watching tennis on TV. John leaves and you increase the volume.	Which sport does he think will be on TV?
	Jim is in a garden labeled vegetables. You pick all the vegetables in the garden and replace them with flowers without Jim’s knowledge.	What does Jim believe is in the garden?
	You and Tim place flowers in the shopping cart. Tim leaves and you ruffle the flowers within the cart.	Where will Tim look for the flowers?
	You and Mary are at the movie theater sharing popcorn out of a container. She leaves and you replace the popcorn with chocolate.	What does she expect to find in the container?
	You placed an apple inside a shoebox while Johnny wasn’t looking. Johnny opens the shoebox.	What does he expect to find inside the shoebox?
	John and you record a tennis match on TV. After recording, John switches the channel to football.	What sport is on the TV recording?
	Ned and you left a car in the street and a bicycle in the garage. While Ned was sleeping, you turned the car on and off.	Where will Ned look for the car?
	An old map shows a rest stop 10 minutes away. The rest stop has since moved further out and is now 30 minutes further away.	How many minutes is the rest stop away?
b	***Additional control tasks***	
	Narrative examples	Question examples
	Self-belief trials	
	You see a jar on a table. You leave the kitchen and the jar falls off the table onto the floor. You return.	Where will you expect to find the jar?
	You put your wallet on the counter and then left the store. The owner looked inside the wallet but kept it on the counter. You return.	Where will you expect to find the wallet?

You place cookies inside a cookie jar. The cookies are replaced with socks without your knowledge. You open the jar.	What do you expect to find inside the jar?
You placed shoes inside a shoebox. The shoes get moved around in the box while you walk to your car. You open the box.	What do you expect to find inside the shoebox?
Second-order belief trials	
Mary and Tom see a jar on a table. Tom leaves the kitchen and Mary moves the jar to the cupboard. Tom returns.	Where does Mary think Tom will look for the jar?
Mary and Tom see a box on a counter. Tom leaves and Mary opens the box while keeping it on the counter. Tom returns.	Where does Mary think Tom will look for the box?
Bob and Maia are drinking hot chocolate. Bob leaves and Maia switches his hot chocolate for apple cider. Bob returns.	What drink does Maia think Bob expects to find in his mug?
Bob and Maia pour lemonade into a glass. Bob leaves and Maia stirs his lemonade. Bob returns.	What drink does Maia think Bob expects to find in his glass?
High inference trials	
Ned and you left a car in the street and a bicycle in the garage. While Ned was sleeping, you switched them.	Where will Ned look for the car?
Ted is browsing his favorite book in the library. Ted leaves and you keep his favorite book in place.	Where will Ted look for his favorite book?
Susy and you see flip flops on a chair and sunglasses on a table. She leaves and you switch the flip flops for glasses. Susy returns.	Where will she look for the flip flops?
You and Tom see a jar on a table. Tom leaves the kitchen and you lift the jar and put it back down. Tom returns.	Where will Tom look for the jar?

Acknowledgments:

M.J is supported by the Banting Foundation, NARSAD Young Investigator Grant, and Foundations of Human Behavior Initiative, B.G. is supported by the NREF and NIH NRSA, E.F is supported by NIH R01DC016607 and R01DC016950, R.B.M is supported by the MGH ECOR and NARSAD Young Investigator Grant, and Z.M.W is supported by NIH R01HD059852, NIH R01NS091390 and the Presidential Early Career Award for Scientists and Engineers.

Data availability:

Details of the participants' demographics and task conditions are provided in Extended Data Tables 1, 2. The behavioral and neuronal data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Wimmer H & Perner J Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13, 103–128 (1983). [PubMed: 6681741]
2. Koster-Hale J & Saxe R Theory of mind: a neural prediction problem. Neuron 79, 836–848 (2013). [PubMed: 24012000]
3. Stone VE, Baron-Cohen S & Knight RT Frontal lobe contributions to theory of mind. J Cogn Neurosci 10, 640–656 (1998). [PubMed: 9802997]
4. Saxe R & Kanwisher N People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage 19, 1835–1842 (2003). [PubMed: 12948738]
5. van Veluw SJ & Chance SA Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. Brain Imaging Behav 8, 24–38 (2014). [PubMed: 24535033]

6. Baron-Cohen S, Jolliffe T, Mortimore C & Robertson M Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome. J Child Psychol Psychiatry 38, 813–822 (1997). [PubMed: 9363580]

7. Brent E, Rios P, Happe F & Charman T Performance of children with autism spectrum disorder on advanced theory of mind tasks. Autism 8, 283–299 (2004). [PubMed: 15358871]

8. Amaral D, Dawson G & Geschwind DH Autism spectrum disorders. (Oxford University Press, New York; 2011).

9. Carruthers P & Smith PK Theories of theories of mind. (Cambridge University Press, Cambridge; New York; 1996).

10. Frith U & Frith CD Development and neurophysiology of mentalizing. Philos Trans R Soc Lond B Biol Sci 358, 459–473 (2003). [PubMed: 12689373]

11. Richardson H, Lisandrelli G, Riobueno-Naylor A & Saxe R Development of the social brain from age three to twelve years. Nat Commun 9, 1027 (2018). [PubMed: 29531321]

12. Williams ZM & Eskandar EN Selective enhancement of associative learning by microstimulation of the anterior caudate. Nat Neurosci 9, 562–568 (2006). [PubMed: 16501567]

13. Saxe R & Powell LJ It's the thought that counts: specific brain regions for one component of theory of mind. Psychol Sci 17, 692–699 (2006). [PubMed: 16913952]

14. Saxe R, Moran JM, Scholz J & Gabrieli J Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. Soc Cogn Affect Neurosci 1, 229–234 (2006). [PubMed: 18985110]

15. Moessnang C et al. Differential responses of the dorsomedial prefrontal cortex and right posterior superior temporal sulcus to spontaneous mentalizing. Hum Brain Mapp 38, 3791–3803 (2017). [PubMed: 28556306]

16. Martin AK, Dzafic I, Ramdave S & Meinzer M Causal evidence for task-specific involvement of the dorsomedial prefrontal cortex in human social cognition. Soc Cogn Affect Neurosci 12, 1209–1218 (2017). [PubMed: 28444345]

17. Dohnel K et al. Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. Neuroimage 60, 1652–1661 (2012). [PubMed: 22300812]

18. Bardi L, Desmet C, Nijhof A, Wiersema JR & Brass M Brain activation for spontaneous and explicit false belief tasks overlaps: new fMRI evidence on belief processing and violation of expectation. Soc Cogn Affect Neurosci 12, 391–400 (2017). [PubMed: 27683425]

19. Fletcher PC et al. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57, 109–128 (1995). [PubMed: 8556839]

20. Apperly IA, Samson D, Chiavarino C, Bickerton WL & Humphreys GW Testing the domain-specificity of a theory of mind deficit in brain-injured patients: evidence for consistent performance on non-verbal, "reality-unknown" false belief and false photograph tasks. Cognition 103, 300–321 (2007). [PubMed: 16781700]

21. Dodell-Feder D, Koster-Hale J, Bedny M & Saxe R fMRI item analysis in a theory of mind task. Neuroimage 55, 705–712 (2011). [PubMed: 21182967]

22. Sabbagh MA & Taylor M Neural correlates of theory-of-mind reasoning: an event-related potential study. Psychol Sci 11, 46–50 (2000). [PubMed: 11228842]

23. Dayan P & Abbott LF Theoretical neuroscience : computational and mathematical modeling of neural systems. (Massachusetts Institute of Technology Press, Cambridge, Mass.; 2001).

24. Arslan B, Verbrugge R, Taatgen N & Hollebrandse B Accelerating the Development of Second-Order False Belief Reasoning: A Training Study With Different Feedback Methods. Child Dev 91, 249–270 (2020). [PubMed: 30474107]

25. Baron-Cohen S, Tager-Flusberg H & Lombardo M Understanding other minds : perspectives from developmental social neuroscience, Edn. Third edition.

26. Bird CM, Castelli F, Malik O, Frith U & Husain M The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. Brain 127, 914–928 (2004). [PubMed: 14998913]

27. Mukamel R, Ekstrom AD, Kaplan J, Iacoboni M & Fried I Single-neuron responses in humans during execution and observation of actions. Curr Biol 20, 750–756 (2010). [PubMed: 20381353]

28. Rizzolatti G, Fadiga L, Gallese V & Fogassi L Premotor cortex and the recognition of motor actions. Brain Res Cogn Brain Res 3, 131–141 (1996). [PubMed: 8713554]

29. Haroush K & Williams ZM Neuronal Prediction of Opponent's Behavior during Cooperative Social Interchange in Primates. Cell (2015).

30. Chang SW, Gariepy JF & Platt ML Neuronal reference frames for social decisions in primate frontal cortex. Nat Neurosci 16, 243–250 (2013). [PubMed: 23263442]

Methods References

31. Williams ZM, Bush G, Rauch SL, Cosgrove GR & Eskandar EN Human anterior cingulate neurons and the integration of monetary reward with motor responses. Nat Neurosci 7, 1370–1375 (2004). [PubMed: 15558064]

32. Patel SR et al. Studying task-related activity of individual neurons in the human brain. Nat Protoc 8, 949–957 (2013). [PubMed: 23598445]

33. Sheth SA et al. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. Nature 488, 218–221 (2012). [PubMed: 22722841]

34. Mian MK et al. Encoding of Rules by Neurons in the Human Dorsolateral Prefrontal Cortex. Cereb Cortex (2012).

35. Erdodi LA et al. Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) processing speed scores as measures of noncredible responding: The third generation of embedded performance validity indicators. Psychol Assess 29, 148–157 (2017). [PubMed: 27124099]

36. Holdnack JA, Xiaobin Z, Larrabee GJ, Millis SR & Salthouse TA Confirmatory factor analysis of the WAIS-IV/WMS-IV. Assessment 18, 178–191 (2011). [PubMed: 21208975]

37. Amirnovin R, Williams ZM, Cosgrove GR & Eskandar EN Experience with microelectrode guided subthalamic nucleus deep brain stimulation. Neurosurgery 58, ONS96–102; discussion ONS196–102 (2006). [PubMed: 16543878]

38. Jamali M et al. Dorsolateral prefrontal neurons mediate subjective decisions and their variation in humans. Nat Neurosci 22, 1010–1020 (2019). [PubMed: 31011224]

39. Nicolelis MAL Methods for neural ensemble recordings, Edn. 2. (Frontiers in Neuroscience, Boca Raton, FL; 2008).

40. Oby ER et al. Extracellular voltage threshold settings can be tuned for optimal encoding of movement and stimulus parameters. J Neural Eng 13, 036009 (2016). [PubMed: 27097901]

41. Perel S et al. Single-unit activity, threshold crossings, and local field potentials in motor cortex differentially encode reach kinematics. J Neurophysiol 114, 1500–1512 (2015). [PubMed: 26133797]

42. Brauner T, Blackburn P & Polyanskaya I Being Deceived: Information Asymmetry in Second-Order False Belief Tasks. Top Cogn Sci 12, 504–534 (2020). [PubMed: 31401814]

43. Shimazaki H & Shinomoto S Kernel bandwidth optimization in spike rate estimation. J Comput Neurosci 29, 171–182 (2010). [PubMed: 19655238]

44. Bowman AW & Azzalini A Applied smoothing techniques for data analysis : the kernel approach with S-Plus illustrations. (Clarendon Press; Oxford University Press, Oxford New York; 1997).

45. Pagan M, Urban LS, Wohl MP & Rust NC Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. Nat Neurosci 16, 1132–1139 (2013). [PubMed: 23792943]

46. Quian Quiroga R, Snyder LH, Batista AP, Cui H & Andersen RA Movement intention is better predicted than attention in the posterior parietal cortex. J Neurosci 26, 3615–3620 (2006). [PubMed: 16571770]

47. Hung CP, Kreiman G, Poggio T & DiCarlo JJ Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863–866 (2005). [PubMed: 16272124]

48. Wasserman L All of statistics: a concise course in statistical inference. (Springer Texts, New York, NY; 2005).

49. Sarafyazd M & Jazayeri M Hierarchical reasoning by neural circuits in the frontal cortex. Science 364 (2019).

50. Cohen R & Elhadad M Syntactic dependency parsers for biomedical-NLP. AMIA Annu Symp Proc 2012, 121–128 (2012). [PubMed: 23304280]

51. Li Z et al. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. BMC Med Inform Decis Mak 19, 22 (2019). [PubMed: 30700301]

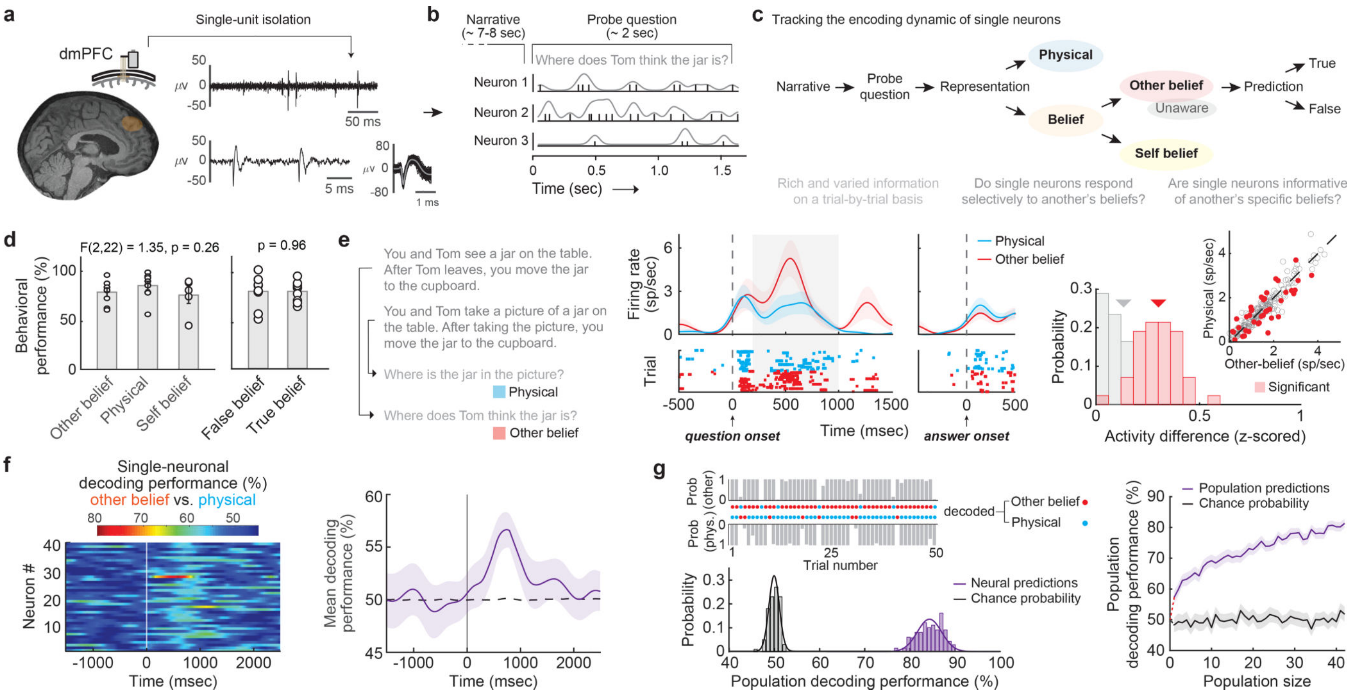


Figure 1. |. Tracking single-cellular representations of another's beliefs in the human dmPFC.

a, Acute single-neuronal recordings were obtained from the superior frontal gyrus of the human dmPFC using microelectrode arrays. **b,** During recordings, the participants performed a verbal variation of the false-belief task. **c,** Schematic illustration of the main task design and trial comparisons (Extended Data Fig. 2 and Extended Data Table 2). **d,** Behavioral performance across the primary task conditions (n =11 participants; self-belief trials were tested in 4 additional controls) represented as mean ± standard error of the mean (s.e.m.). **e,** Representative narrative and question examples for other-belief vs. physical trials are shown to the *left*. In the *middle* is a peri-stimulus time histogram (± s.e.m.) and spike raster reflecting the activity of a representative neuron during questioning. On the *right*, the firing rates (*inset*) and z-scored activities for neurons with (n = 42) and without (n = 170) significant modulation are displayed in red and gray, respectively. **f,** A linear discriminant quantified the degree to which the activities of individual neurons (*left*, n = 42) were predictive of other-belief vs. physical trials on a trial-by-trial basis (one-sided permutation test, $p < 0.025$). The time course of mean decoding performance with 95% confidence limits (CL) is shown on the *right*. **g,** Decoding projections for individual trials as well as decoding performance for the neural population (*left, n* = 42) and its cumulative (*right,* mean with 95% CL) are displayed. The p-values by one-way ANOVA and two-sided unpaired t-test in **d** (*left* and *right* panels, respectively) are indicated.

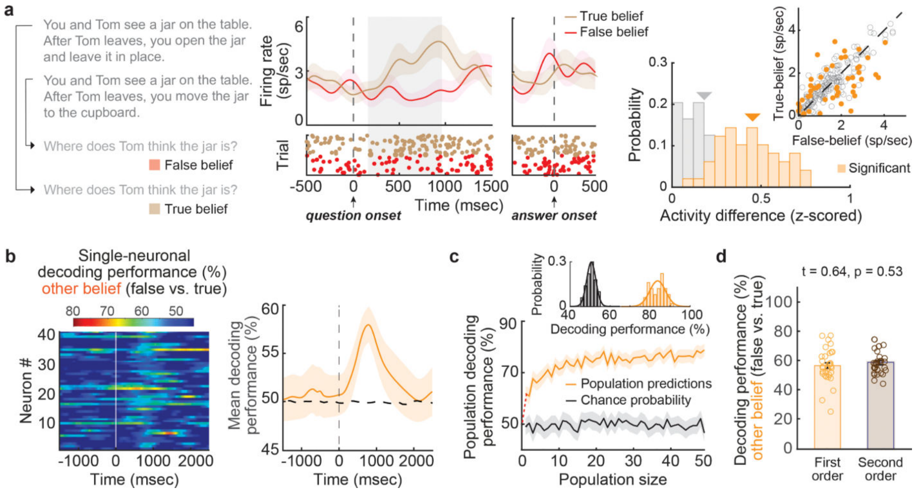


Figure 2. |. Single-neuronal predictions of another's true and false beliefs.

a, Representative narrative and question examples for false- vs. true-belief trials are shown to the *left*. In the *middle* is a peri-stimulus time histogram (± s.e.m.) and raster reflecting the spiking activity of a representative neuron during questioning. On the *right*, the firing rates (*inset*) and *z*-scored activities for neurons that displayed significant modulation ($n = 49$) are in orange and those that did not ($n = 163$) are in gray. **b,** A linear discriminant quantified the degree to which the activities of individual neurons were predictive of false-belief vs. true-belief trials on a trial-by-trial basis (*left*, $n = 49$; one-sided permutation test, $p < 0.025$). The *right* traces illustrate the time course of the mean decoding performance with 95% CL. **c,** Mean and cumulative decoding performances for the population of neurons ($n = 49$) with their 95% CL. **d,** Single-neuronal decoding performances (mean ± s.e.m., $n = 32$ neurons) during first-order and second-order false- vs. true-belief trials ($n = 4$ participants) were similar (two-sided paired *t*-test, $p = 0.53$). Similar results were obtained when considering mean firing rates (1.47 ± 0.12 vs. 1.49 ± 0.12 spike/sec; two-sided rank-sum test: *z*-value = 0.05, $p = 0.96$), respectively.

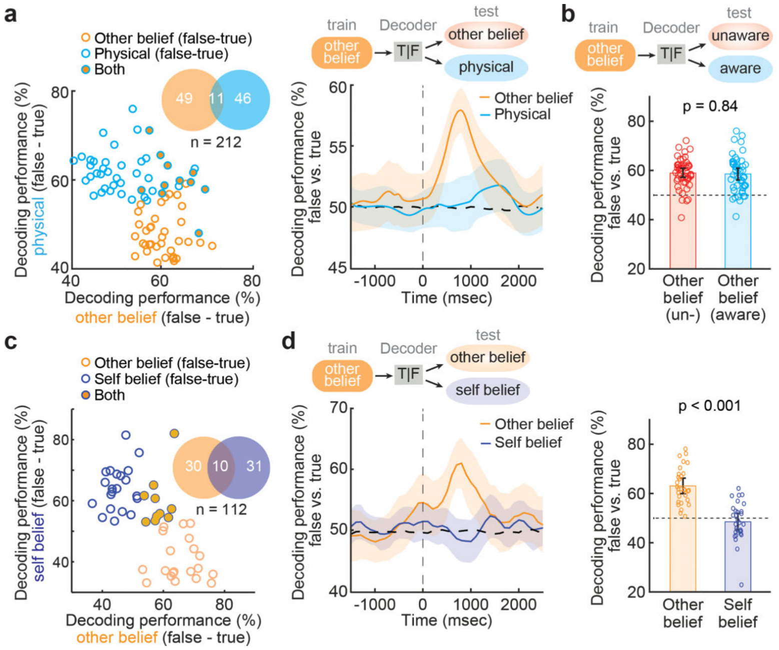


Figure 3. |. Neuronal responses to self vs. others' beliefs and perspective.

a, The scatter plot illustrates decoding accuracies for each cell comparing false vs. true other-belief and false vs. true physical trials. The lack of significant overlap indicates that most other-belief encoding neurons encoded no information about the physical state of reality (one-sided permutation test, $p = 0.36$). On the *right*, decoding performances (mean with 95% CL, orange) for these neurons ($n = 49$) dropped to chance level when the same neurons were used to decode true- vs. false-physical trials (blue). **b,** Similar decoding performances (mean with 95% CL) were observed across all false vs. true other-belief neurons ($n = 49$) under situations in which the social agent was aware or unaware of events (two-sided paired *t*-test, $p = 0.84$, see Extended Data Fig. 6) suggesting that they reliably tracked the other's perspective **c,** The scatter plot illustrates decoding accuracies for each cell comparing false vs. true beliefs for self or other. The lack of significant overlap indicates that most other-belief encoding neurons encoded no information about the participant's own imagined beliefs (one-sided permutation test, $p = 0.14$). **d,** The time course (*left*) and the corresponding individual neuronal (*right*) decoding performances (mean with 95% CL) are shown for neurons ($n = 30$) that predicted whether the other's beliefs were true vs. false (orange) as well as when the same neurons were used to decode true vs. false self-belief trials (purple).

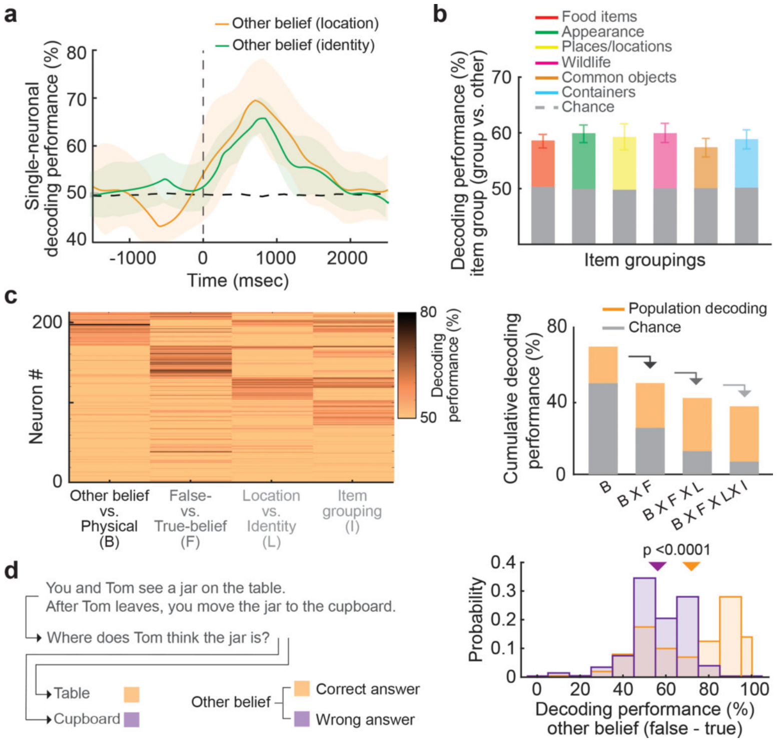


Figure 4. |. Population predictions of another's belief contents and their relation to behavioral performance.

a, Mean decoding performance with 95% CL for neurons that accurately differentiated between false- vs. true-belief trials ($n = 49$) based on whether the social agent's beliefs involved an item's identity or its location. b, The bar graphs indicate the accuracies (mean with 95% CL) with which the neurons could predict the specific item grouping being considered by the social agents from all other possible groupings ($n = 200$ repetitions). c, On the *left* are mixed population predictions sorted based on the four primary features that the participant had to consider in order to correctly infer the other's beliefs. On the *right* are the summative decoding performances for the mixed population. d, The histograms demonstrate the decoding accuracies for false vs. true other-belief neurons on trials in which the participants provided the correct vs. incorrect answers (Extended Data Fig. 8). The arrows indicate mean decoding performances (one-sided permutation test, $p < 0.0001$).