Neural Correlates of Self-Referential Belief Processes

Emily Bruns[1,2,3], Immanuel Scholz[1,2], Georgia Koppe[3,4,5,6], Peter Kirsch[1,2,7], & Martin Fungisai Gerchen[1,2,3,7]

[1]Department of Clinical Psychology, Central Institute of Mental Health, University of Heidelberg/Medical Faculty Mannheim, Mannheim, Germany

[2]Department of Psychology, University of Heidelberg, Heidelberg, Germany

[3]Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany

[4]Department of Theoretical Neuroscience, Central Institute of Mental Health, University of Heidelberg/Medical Faculty Mannheim, Mannheim, Germany

[5]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, University of Heidelberg/Medical Faculty Mannheim, Mannheim, Germany

[6]Hector Institute for AI in Psychiatry, Central Institute of Mental Health, Mannheim, Germany

[7]Bernstein Center for Computational Neuroscience Heidelberg/Mannheim, Mannheim, Germany

Abstract

Background: Belief processing as well as self-referential processing have both been consistently associated with cortical midline structures. In addition, seminal neuroimaging papers have implicated cortical regions such as the vmPFC in *general* belief processing. However, the neural correlates of *self-referential* belief are yet to be investigated in functional magnetic resonance imaging (fMRI).

Methods: In this fMRI study, we presented 120 statements with trait adjectives as target words to N=27 young healthy participants and asked them to judge whether they believed that these trait adjectives applied to themselves, a self-chosen close person, or a public person (the German chancellor at that time). Participants subsequently rated how certain (0-100%) they were in their judgment.

Results: As expected, self-referential processing evoked a large cluster in the vmPFC, ACC and dmPFC. For belief, we found an activated cluster in the vmPFC during statement presentation, which partly overlapped with the cluster for self-referential processing. The cluster for self-belief vs. disbelief was similar in location and size to the cluster for general belief processing and distinct from the cluster for self-referential processing. We also found dmPFC activation for uncertainty in belief evaluations.

Discussion: We successfully replicated vmPFC involvement in belief processing and found a common neural correlate for belief and self-belief in the vmPFC. The activation clusters for self-belief versus self-referential processing were distinct, implying distinct neural processes. This insight will prove relevant for investigations in clinical populations with aberrant (self-)belief processing. Furthermore, we replicated the role of the dmPFC in uncertainty, supporting a dual neural process model of belief and certainty.

Introduction

Belief is a substantial aspect of everyday human experience and crucial in shaping subjective reality, emotion, and behavior (Harris et al., 2008; James, 1890/1910). It has far-reaching implications for the public and political domain, including societal polarization and the spread of conspiracy theories (Bowen et al., 2023; Gerchen et al., 2023; Haghtalab et al., 2021), to inflexible individual convictions in mental disorders like delusion and depression.

While belief has not been a dominant focus of psychological research in the last decades, it is a traditional psychological construct. William James conceptualized belief as "the mental state or function of cognizing reality" and "a sort of feeling more allied to the emotions than to anything else" (1890/1910; p. 283), highlighting its emotive and subjective component. The nature of belief processes has been intensively discussed in the early psychological literature. Already in 1859, Alexander Bain postulated that the opposite of belief is doubt or uncertainty: "The real opposite of belief as a state of mind is not disbelief, but doubt, uncertainty; and the close alliance between this and the emotion of fear is stamped on every language." (Bain, 1859, p. 574), introducing the dimension of certainty and highlighting its importance for an individual's emotional state. The more recent psychological literature particularly focuses on the relationship of belief and disbelief processes (Asp 2012; Asp, 2022; Gilbert, 1991; Gilbert, 1993; Hasson, 2005).

Nowadays, neuroimaging approaches are one of the most promising ways to further elucidate the processes underlying belief (see for example Harris et al., 2008; Harris et al., 2009; Marques et al., 2009; Douglas et al., 2013; Seitz, 2022; Seitz & Angel, 2012; Seitz et al., 2018). In a seminal fMRI study by Harris et al. (2008), participants were shown statements from different categories (autobiographical, mathematical, geographical, religious, ethical, semantic, and factual), and judged whether they were "true", "false", or "undecidable". In this study, the authors associated the ventromedial prefrontal cortex (vmPFC) with belief. The vmPFC has been linked to behavioral modulation to rewards and goals (Grabenhorst & Rolls, 2011; Hornak et al., 2004; Matsumoto & Tanaka 2004), emotional saliency (Goel & Dolan, 2003) as well as the recruitment of somatic markers in decision-making (Bechara et al., 2000; Damasio, 1996). Notably, in the context of belief processing, the vmPFC has also been proposed as a neural correlate of a "false tagging" process, phenomenally manifesting as doubt and skepticism (Asp et al., 2012). Interestingly, this notion is consistent with a larger deflection of vmPFC activity for disbelief than for belief, which creates the positive contrast effect for belief (Harris et al., 2008; Gerchen et al., 2023; see also Supplement, Figure 1). For disbelief, Harris et al. (2008) found effects in the left inferior frontal gyrus, anterior insula, dorsal anterior cingulate (dACC), superior frontal gyrus, and superior parietal lobule. An important limitation of the study by Harris et al is, however, that belief was counfounded with (un)certainty, since the authors did not assess certainty in "true" and "false" belief evaluations, or uncertainty in "undecidable" statements.

Based on Harris et al.'s (2008) paradigm, Gerchen et al. (2023) conducted an fMRI study to investigate general belief in factual, political, religious, conspiracy, and superstitious, but not in self-referential, statements. In contrast to Harris et al. (2008), we (Gerchen et al. 2023) temporally separated statement presentation from truth evaluation and applied a forced choice decision "true"/ "false" decision with a separate certainty assessment. Interestingly,

we found similar vmPFC activation for belief as Harris et al. (2008), but at the time of judgment and not during the preceding statement presentation. During statement presentation, we identified activity in bilateral dorsolateral prefrontal cortex (dlPFC), left superior parietal cortex, and left lateral frontopolar cortex (left lFPC) for statements evaluated as "true". For disbelief, we found activation in the temporal pole (TP), a structure serving as an amodal semantic hub (Chen et al., 2017; Patterson et al., 2007). The TP is also involved in the processing of semantic meaning (Tsapkini et al., 2011; Vandenberghe et al., 2002) and semantic memory (Bonner & Price, 2013; Simmons & Martin, 2009), and has been implicated in belief-laden as opposed to belief-neutral reasoning (Goel & Dolan, 2003). Notably, the strongest effect we found was activation in the dorsomedial prefrontal cortex (dmPFC) associated with uncertainty, which we interpreted as a marker of the "subjective epistemic risk" of a statement. Based on this finding, we propose dual neural belief processes with veracity/falsehood and uncertainty/certainty as two distinguishable factors.

An important further question is whether belief processes are general processes or whether different functions are involved in belief in different domains. For example, comparing religious and non-religious belief, Harris et al. (2009) found that religious belief had higher associations with brain processes involved in emotion, self-representation, and cognitive conflict. In another study, Howlett & Paulus (2015) found that testability of statements was associated with activation in the dlPFC and posterior cingulate cortex.

One of the most relevant potential distinctions of belief domains in humans is the differentiation between self-referential and non-self-referential belief. Better understanding of self-referential belief processing is especially important for laying the foundation for future studies in clinical contexts, as self-beliefs and aberrant self-referential processing are highly relevant in mental disorders. In fact, negative self-beliefs are associated with mental health issues across disorders (e.g., depression, psychosis, eating disorders) and cultures and ethnic groups (Cooper & Cowen, 2009; Dozois & Rnic, 2015; Nguyen et al., 2019; Orth & Robbins, 2013; Smith et al., 2006). For example, self-referential cognition is a pivotal component of Beck's cognitive triad of depression (1967), and aberrant beliefs and belief updating are hallmarks of psychosis (Fletcher & Frith, 2009; Kube & Rosenkrantz, 2021). It is thus imperative that future clinical studies on the diagnosis and treatment (for psychosis see Pott & Schillbach, 2022) of these disorders consider how changes in belief processing are represented at the neural level.

For the first study aiming to make this differentiation, Han et al. (2017) developed an fMRI paradigm based on trait adjectives describing the participant or a gender-matched celebrity. However, they did not focus on the differentiation between self-referential and non-self-referential belief, but rather on the contrast of "believing" vs. "thinking". For this contrast, they found stronger activation in the left anterior insula (AI) / inferior frontal cortex (IFC) as well as greater intrinsic connectivity between the left AI / IFC and the left occipital cortex. Regarding self-belief, they found that, in the belief condition, self-judgments were associated with greater activation in the left AI / IFC as well as greater functional connectivity between the mPFC and the left occipital cortex than valence-judgments during judgment. Using the same paradigm, Gao et al. (2022) compared neural belief processes between Chinese and Danish participants. For both cultural groups, they identified the mPFC in association with general belief over both conditions. In an analysis secondary to their main research question, they found that self- vs. celebrity-related belief was associated with activation in the vmPFC.

The finding by Gao et al. (2022) of a stronger self-referential belief effect in the vmPFC emphasizes the question whether self-referential and non-self-referential belief processes are based on similar or distinct neural processes. While belief processing as well as self-referential processing (Kim, 2012; Kurczek et al., 2015; Lemogne et al., 2011; Northoff et al., 2006) have both been associated with cortical midline structures (CMS) overlapping in the vmPFC, further and more explicit investigation is warranted as to how this rough correspondence is reflected in actual activations, and as to whether the involved CMS differ for belief of self-referential content.

In the present study, our aim was therefore to investigate the neural correlates of self-referential belief in comparison to non-self-referential belief. To address this question, we designed a paradigm related to Han et al. (2017). In our paradigm, participants judged short statements about themselves, a close person (a self-chosen close friend or family member), and a public person (the long-time German chancellor at the time of the experiment, Angela Merkel). As in our former work (Gerchen et al., 2023), responses were given as a binary forced-choice belief decision followed by a certainty rating for the decision (Figure 1). This approach also allowed us to compare the obtained results for self-referential belief to those of Gerchen et al. (2023) for general belief and investigate differences as well as the replication of effects.

Materials & Methods

Participants

In total, N=32 young healthy participants were recruited for the study. The participants were mainly university students from the German cities Mannheim, Heidelberg, and Freiburg. All participants were screened by telephone for exclusion criteria prior to participation. Exclusion criteria were left-handedness, uncorrected impaired vision, a diagnosis of a neurological or mental disorder, pregnancy, implants or metal parts in the body, and claustrophobia. The data sets of five participants had to be excluded due to anatomical incidental findings (1), inadequate language skills (1), and excessive head motion (3). Thus, N=27 participants (14 women; average age 23.3 years; age range 19-38 years) were included in the analyses. All participants were fully informed about the study and provided written informed consent. After the experiment, participants received either a reimbursement of 30€ or partial course credit at their university. The study was approved by the ethics committee of the Medical Faculty Mannheim, Heidelberg University, Germany (2020-649N) and procedures complied with the Declaration of Helsinki of the World Medical Association.

Data Acquisition

MRI measurements were conducted on a Siemens Magnetom TrioTIM 3T scanner (Siemens Healthineers, Erlangen, Germany) at the Central Institute of Mental Health in Mannheim, Germany. Data for each participant was acquired in a single session that consisted of an anatomical measurement, a 6-minute resting state scan, and the task. Anatomical images were acquired with a magnetization prepared rapid gradient echo (MPRAGE) sequence with repetition time TR=2.3s, echo time TE=3.03ms, flip angle=9° and an isotropic resolution of 1x1x1mm. Functional images were acquired with an Echo Planar Imaging (EPI) sequence with TR=1.64s, TE=30ms, flip angle=73° and GRAPPA factor 2 in 30 slices of 3mm thickness with 1mm gap, an in-plane resolution of 3x3mm, and field of view FoV=192mm. In addition, respiratory and cardiac activity was recorded during functional measurements with built-in equipment.

Experimental Task

The experiment consisted of three different conditions: belief about the self, a self-chosen close friend or family member ("close person"), and the previous German chancellor, Angela Merkel ("public person"). Two other-conditions were used to prevent a mono-operation bias (Cole et al., 1981) and to assess potential effects of intimacy and autobiographical processing. Angela Merkel was chosen because she is a well-known public figure, whom we did not expect the participants to know personally (which none of the participants did, according to their self-report). All three conditions aimed to trigger subjective belief processing of personality traits (e.g., "I am / [Name of close person] is / Merkel is lazy."). The experiment used a rapid serial visual presentation paradigm. First, a fixation cross appeared on a monitor for 1 s, followed by three individual words each shown for 300 ms. The final target word, which conveyed the meaning of the statement, was presented for a jittered period of 3.1 to 4.56 s. Subsequently, the participants were prompted to a self-paced belief evaluation of the statement ("This statement is true."; German: "Diese Aussage ist wahr."). This prompt was included, as Wiswede et al. (2013) postulated truth evaluation to

depend on task demands and an evaluative mindset rather than being an automatic process. The "yes" or "no" answers were randomly assigned to the left or right button of a small hand-held four-button controller to prevent confounding preparatory motor activation during the evaluation period. Afterwards, participants rated how certain they were in their evaluation ("I am ... percent certain."; German: "Ich bin mir sicher zu ...") on a visual analogue scale from 0% to 100% in their own time. The start value was set to 50% and participants could navigate to their certainty value of choice using the hand-held controller. An exemplary trial is shown in Figure 1. Participants completed 120 trials in the fMRI scanner. The average duration of the task was 26:13 ± 2:52 min (range: 23:14-37:01 min).

A pool of trait adjectives was generated specifically for this paradigm because there were no preexisting trait-adjective lists normed for endorsement of the trait adjective for the self, a close person, and a public figure. We aimed for an even distribution of statements rated as true and false, and thus included negative, neutral, and positive trait adjectives. The final 80 items drew on a pool of trait adjectives from Lazzari et al. (1978) and King (1983), of likability-rated items from Dumas et al. (2002), Schonbach (1972) and Alicke (1985), and on self-generated trait adjectives.

We designed our experiment to be conducted with two parallel statement sets subsequently with magnetoencephalography (MEG) and fMRI. To split the statements into two sets, 9 participants were recruited to pre-test the paradigm including all 80 target words. Based on the pre-test data, the pool of target words was partitioned into two maximally equivalent and internally heterogeneous sets of 40 items for the parallel test versions using Papenberg and Klau's (2021) anti-clustering method. The parallel versions were used for MEG and fMRI and counterbalanced over participants. The order of items within a set at presentation was random. In this paper, only the fMRI data is presented.

Data Analysis

Data analyses were conducted in MATLAB (R2017a; The MathWorks, Inc., Natick, Massachusetts, USA). fMRI data was analyzed with SPM12 (v7219; Wellcome Center for Human Neuroimaging, London, United Kingdom).

fMRI Preprocessing

The anatomical images were segmented and normalized to the SPM12 TPM MNI template. Functional images were slice-time corrected, realigned to the mean image, co-registered to the anatomical image and normalized to MNI space by applying the forward deformation matrix estimated from the anatomical image. In this step, the functional data was rescaled to 3x3x3mm isotropic resolution and subsequently smoothed with a three-dimensional Gaussian kernel with full width at half maximum FWHM=8x8x8mm. In addition, nuisance signals from regions containing white matter (WM) and cerebrospinal fluid (CSF) were estimated, the ART toolbox (http://www.nitrc.org/projects/artifact_detect) was used to identify volumes affected by movements (framewise displacement FD>0.5mm; scan-to-scan global signal change z>4), and the TAPAS PhysIO toolbox (Kasper et al., 2017) was used to estimate physiological nuisance regressors based on the acquired cardiac and respiratory signals.

fMRI First Level Analyses

fMRI first level analyses were conducted using general linear models (GLMs). The main first level model to estimate self-referential and belief processes contained 6 experimental regressors of statement presentations (belief and disbelief for the self, close person, and public person conditions, respectively) convolved with the canonical hemodynamic response function (cHRF). Based on this model, the contrasts 'self vs. other' (other including the close and public persons) over belief and disbelief statements, and 'belief vs. disbelief' over all persons as well as within the conditions were estimated. The first level model for estimating uncertainty effects contained three experimental regressors for statement presentations in the conditions (self, close person, public person) convolved with the cHRF as well as their parametric modulation by trial-wise certainty ratings. Uncertainty effects were estimated by negative contrasts on the parametric modulation regressors over all conditions as well as within the conditions.
Both first-level design matrices contained regressors of the time courses of fixation cross presentations and button presses, and the decision and certainty phase convolved with the cHRF, as well as six motion regressors, WM and CSF regressors, dummy regressors for movement-affected time points, and the physiological nuisance regressors.

fMRI Second Level Analyses

Participant-specific contrast maps from the first level analyses were entered into second level analyses with one-sample t-tests. All second level models contained sex as covariate. A significance threshold of p=0.05 cluster-level FWE-corrected with cluster-defining threshold (CDT) of p=0.001 uncorrected was applied in all second level analyses.

Overlap with Gerchen et al. (2023)

For descriptive replication analyses, we plotted the overlap of the identified activations with the 'belief > disbelief' effect in the vmPFC and the uncertainty effect in the dmPFC we previously identified in a separate sample with belief judgements on statements about facts, politics, religion, conspiracy theories, and superstition (Gerchen et al., 2023). Here, we applied to the data of Gerchen et al. (2023) the original reported threshold for the vmPFC effect (p=0.05 cluster-level FWE significant with CDT=0.001 unc.) and a stricter than originally reported significance threshold for the dmPFC effect (p=0.05 whole-brain FWE corr.) for illustrative reasons.

Results

Behavior

On average, 60.56 ± 7.11% (inter-subject range 44-72%) of statements were judged as true. The self-referential condition showed the highest (65.3 ± 9.56%) and the public-person condition the lowest (53.47 ± 7.64%) rate of belief (paired-sample t-test t(26) = 6.58, p<0.001, g=1.23). 62.89 ± 9.08% of statements about the close person were believed to be true, revealing a significant difference only to the public person (comparison paired-sample t-test t(26) = -4.93, p<0.001, g=-0.92) but not the self (paired-sample t-test t(26) = 1.6, p=0.1221, g=0.3), Participants were 72.99 ± 8.07% (range 56-86%) certain in their belief evaluation on average. They were most certain judging their chosen close person (79.36 ± 8.85%) and least certain judging the public person (62.14 ± 14.68%; comparison paired-sample t-test t(26) = 5.91, p<0.001, g=1.10). On average, participants were 77.4% ± 8.51% certain in belief evaluations about themselves, which was significantly higher than certainty in evaluations about the public (t(26)=4.93, p<0.001, g=0.92 but not the close person (t(26) = -2.01, p = 0.0544, g = -0.38)). Overall, certainty was significantly higher in statements rated as true (75.09 ± 8.47%) than in those believed to be false (69.83 ± 8.42%; paired-sample t-test t(26) = 5.59, p<0.001, g=1.04).

When we compared the two item sets, they did not differ in the rate of belief in the self-referential and close person conditions. However, the rate of belief differed in the public person condition. For certainty, there were no differences between the sets in any of the three conditions (see Supplementary Table 3).

fMRI self vs. other

Testing the contrast 'self > other' (other including the close and public persons) resulted in the largest cluster of activation found in this experiment. Namely, we identified a significant cluster centered in the vmPFC and ACC, which further reached into the dmPFC, corpus callosum and subcortical structures in the diencephalon (Figure 2).

fMRI belief vs. disbelief

For the contrast 'belief > disbelief', we found a significant cluster in the ACC and vmPFC across all conditions (Figure 3a). As in Gerchen et al. (2023), this 'belief' effect was based on stronger deactivation of the vmPFC cluster in the disbelief condition compared to the belief condition (Supplementary, Figure 1), but was present during the presentation phase of the experiment. While we did not find significant differences for the contrast 'belief > disbelief' between conditions, the within-condition contrast 'belief > disbelief' was only significant in the self-referential condition but not for the close or public persons. The cluster identified in this contrast ('self belief > self disbelief') involves the ACC, vmPFC and a small part of the genu of the corpus callosum and is very similar to the cluster for general 'belief > disbelief' in shape and localization (Figure 3b). Moreover, it partly overlapped at the edges with the activation for the contrast 'self > other' shown in Figure 2 but was distinct from it in size and extent (see Supplementary Figure 3). Exploratively, small subcortical clusters were found in the amygdala and ventral striatum (n.s., cluster-level p=0.09; see Table 2). The activation for self-referential belief showed substantial overlap with the findings on general belief from the fMRI study by Gerchen et al. (2023) and involved the same cortical regions

(ACC, vmPFC; see Figure 5a). The contrast 'disbelief > belief' revealed no significant clusters.

fMRI uncertainty

We identified a cluster in the dmPFC for uncertainty across all conditions (Figure 4a). Remarkably, our cluster overlapped with the activation identified for uncertainty in Gerchen et al. (2023) (Figure 5b). Again, a within-condition uncertainty effect could only be found in the self condition (Figure 4b) but not for the close or public persons. The cluster for uncertainty in self-referential belief evaluations looked different from the cluster for uncertainty across all conditions in shape but was also localized in the dmPFC. Findings for the opposite contrast direction (certainty) that were also present are shown in the supplement (Supplementary, Figure 2).

Discussion

In the current study, we investigated the question whether the neural correlates of self-referential belief processing differ from those of general belief. For this purpose, we had participants evaluate their belief in statements about themselves, a close, and a public person. While we found belief effects in the vmPFC (Figure 3a) and uncertainty effects in the dmPFC over all conditions (Figure 4a) and within the self-referential condition (Figure 4b), we did not find differences between the conditions, and the effects identified in the self-referential condition were overlapping with our previous effects for general belief processing (Figure 5; Gerchen et al., 2023).

With respect to behavior, in our study the distribution of true versus false statements was slightly skewed across conditions, indicating marginally unbalanced item endorsement. The distribution of statements rated as true versus false also varied between the conditions, with the highest rate of endorsement in the self condition and the lowest in the public person condition. This may be explained by item selection as well as choice of the public person, i.e. the previous German chancellor, who is a Conservative politician while most student participants (and their close peers) might be more progressive.
The certainty ratings also differed between conditions. On average, participants were most certain in their belief evaluations regarding the self-chosen close person and least certain for the public person. The finding that participants placed greater certainty in self-referential judgments than in the judgments of a person who is not a friend or family member is in line with an early report by Kuiper and Rogers (1979). The relatively low certainty in truth evaluations of the public person might be explained by the lack of personal connection and shared episodic memories.
Notably, certainty in belief evaluation was significantly greater than certainty in disbelief evaluations. Interestingly, this finding is inconsistent with the results of Gerchen et al. (2023), where the effect was inverse, which might be due to the different nature of statements used. It might be the case that participants more vehemently disagreed with superstitious or conspiracy statements used in Gerchen et al. (2023), while they were more uncertain in abstract evaluations of themselves or the close and other persons.

As expected, self-referential processing (self > other) revealed a large cluster in the vmPFC and ACC, also reaching into the dmPFC, corpus callosum and subcortical structures (Figure 2).

Belief vs. disbelief evaluation was associated with activation in the vmPFC and ACC (Figure 3a). Within conditions, the contrast 'belief > disbelief' was only significant in the self-condition (Figure 3b), with the cluster looking almost identical to the one found over all conditions in location, size, and shape. The lack of significance in the close and public person conditions in our study might be due to low power and stronger neural activation for the self, as target words describing the self rather than another person are salient and self-relevant stimuli (Sui & Humphreys, 2015). This would also be consistent with Gao et al. (2022), who found significantly greater vmPFC activation for self-belief than belief in statements about a gender-matched celebrity. It is, however, important to note that in our case we did not identify a significant difference between the conditions.

Interestingly, the cluster identified for self-referential belief was in close vicinity to the one identified for self-referential processing (self > other) but differed in exact location, size, and shape (Supplementary Figure 3). These distinguishable clusters imply that self-referential processing and belief processing are *distinct* neural processes embedded in the same neural networks, bearing important implications for later investigations in clinical populations displaying aberrant belief and/or self-referential processing.

Importantly, the cluster we identified for self-referential belief replicates the vmPFC belief processing effect in Gerchen et al. (2023; Figure 5a, Supplementary Figure 1). Notably, Gerchen et al. (2023) identified the effect during the decision phase and linked it to decision making, whereas here it was identified during the statement presentation phase. This difference might be related to the higher temporal precision and the simpler statement structure (3 words in comparison to complex statements) in the present study, which could lead to faster and more precise decision processes. Nonetheless, future research is needed to clarify the exact role of the vmPFC in belief processing. In the context of believing and deciding, the vmPFC appears to play a crucial role in belief processing of statements about the self, others, and a variety of other categories (e.g., Harris et al., 2008; Gao et al., 2022; Gerchen et al., 2023), "false tagging" dubious cognitive representations (Asp et al., 2012), signaling emotional saliency (Goel & Dolan, 2003) as well as recruiting somatic markers for decision-making (Bechara et al., 2000; Damasio, 1996).

In sum, our findings suggest a *common neural correlate* for self-referential belief and general belief processing in the vmPFC. As Gao et al. (2022) and our descriptive results show, the vmPFC effect might be even slightly stronger in self-referential processing, but the localization of the effect is clearly the same.

In contrast to Gerchen et al. (2023), we did not identify the left lPFC, left superior parietal cortex or dlPFC for belief processing during statement presentation. Gerchen et al. (2023) suggested that the lPFC's role in relational integration of working memory contents might lead to its proposed role in generating a belief signal. A reason why we did not find the dlPFC or superior parietal cortex in this study might be the lower working memory load (role of dlPFC in working memory: e.g., Barbey et al., 2013; role of superior parietal cortex in working memory: e.g., Koenigs et al., 2009), as participants only had to retain the target word and person to be evaluated, whereas Gerchen et al. (2023) used more complex statements. On the other hand, the lPFC has been shown to be involved in autobiographical memory search (Cabeza & St Jacques, 2007). It remains thus unclear why it was not identified in this study.

We also did not identify any region for opposite contrast 'disbelief > belief', while Gerchen et al. (2023) found the anterior TP. This might be a result of different experimental designs and low power and should be further investigated in future studies. A plausible explanation as to why we did not replicate the anterior TP might be because, contrary to Gerchen et al. (2023), we did not include factual statements of semantic categories and the TP has repeatedly been implicated in semantic processing and memory (Bonner & Price, 2013; Chen et al., 2017; Patterson et al., 2007; Simmons & Martin, 2009; Tsapkini et al., 2011; Vandenberghe et al., 2002).

With respect to the uncertainty/certainty factor of belief postulated by Gerchen et al. (2023), we identified a cluster in the dmPFC associated with uncertainty that was again present over all conditions (Figure 4a) and within the self-referential condition (Figure 4b). Notably, this effect replicates the findings of Gerchen et al. (2023) we interpreted as a "subjective epistemic risk" marker despite the different nature of statements, again implying a *common neural correlate* of self-referential and general belief for this process.

In another context, Kaplan et al. (2016) found that dmPFC activation during challenges to political positions negatively correlated with belief change, which is coherent with its supposed role in uncertainty. This process might thus become relevant in future studies in clinical populations with heightened risk processing and uncertainty (e.g., eating disorders, Brown et al., 2017; anxiety, depressive and obsessive-compulsive disorders; Gentes & Ruscio, 2011), where dmPFC activation might serve as a diagnostic marker or indicator of change processes.

Notably, several limitations apply to our study. First, the sample mainly consisted of university students and is thus not representative of the general population. Second, the analyses were likely underpowered, as we only attained a final sample size of 27 participants. This might play a part in why we did not find significant effects for the contrast 'belief > disbelief' within the close and public person conditions, as activation in these regions might be weaker for other- than for self-reference as discussed above. Third, participants might have been less vigilant and conscientious at the time of measurement in the fMRI due to the previous run of the MEG part of the experiment. Fourth, the two parallel sets of target items were not perfectly balanced for the public person statements, which might be due to the limited sample of N=9 for the pre-test. Finally, the distribution of true vs. false and certainty ratings was skewed between the three conditions. Thus, further future well-powered and confirmatory analyses are warranted to corroborate our findings.

Conclusion

By demonstrating common neural correlates of self-referential and general belief processes in the vmPFC and dmPFC and distinguishing self-referential processing from self-referential belief processing, our findings lay a foundation stone for future studies on belief processing. Further investigating these processes might be especially relevant in clinical populations with aberrant (self-referential) belief (e.g., depression, psychosis) and certainty processing (e.g., anxiety, depressive, eating, and obsessive-compulsive disorders), where they might serve as a diagnostic marker and indicator of intervention related change processes.

Figures & Tables
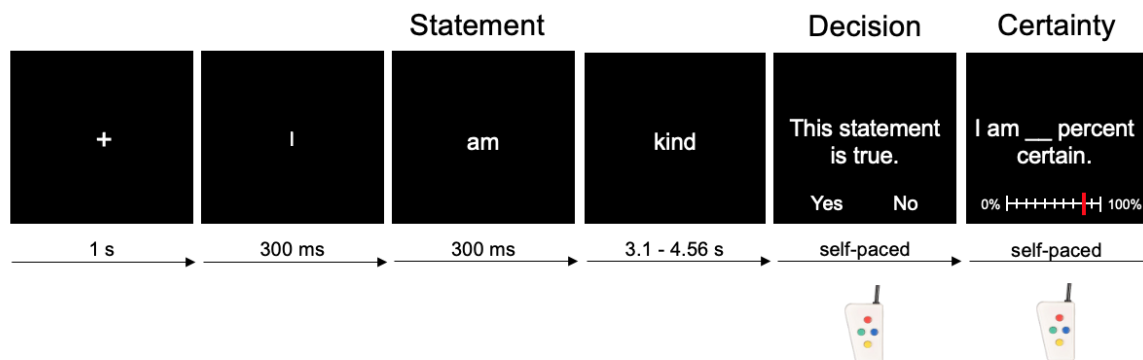
Figure 1. Statement Judgment Task. 120 statements containing trait adjectives as target words were presented in random order. The target word was presented for a jittered duration of 3.1-4.56 s. Participants were not required to react during the rapid visual presentation but were subsequently prompted to make a forced-choice, self-paced binary decision about the truth / falsehood of that statement. Thereafter, participants evaluated their certainty in the previous belief evaluation on a visual analog scale from 0-100% in their own time. The original statements were presented in German.
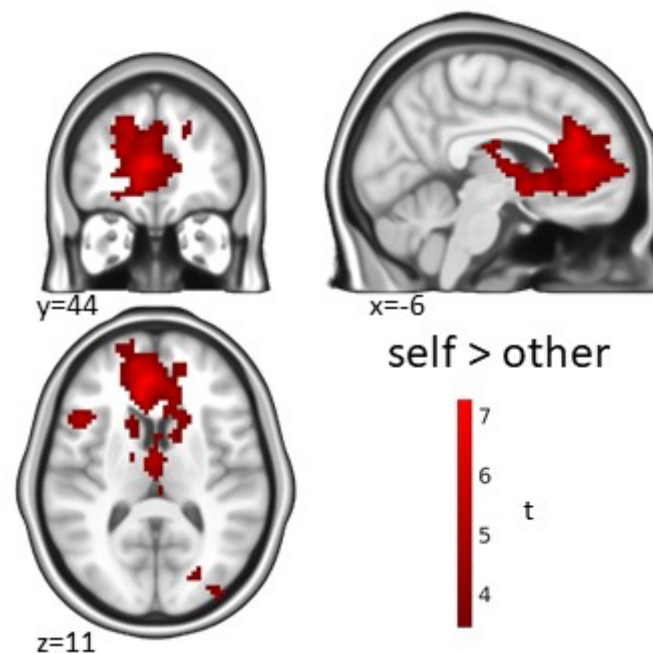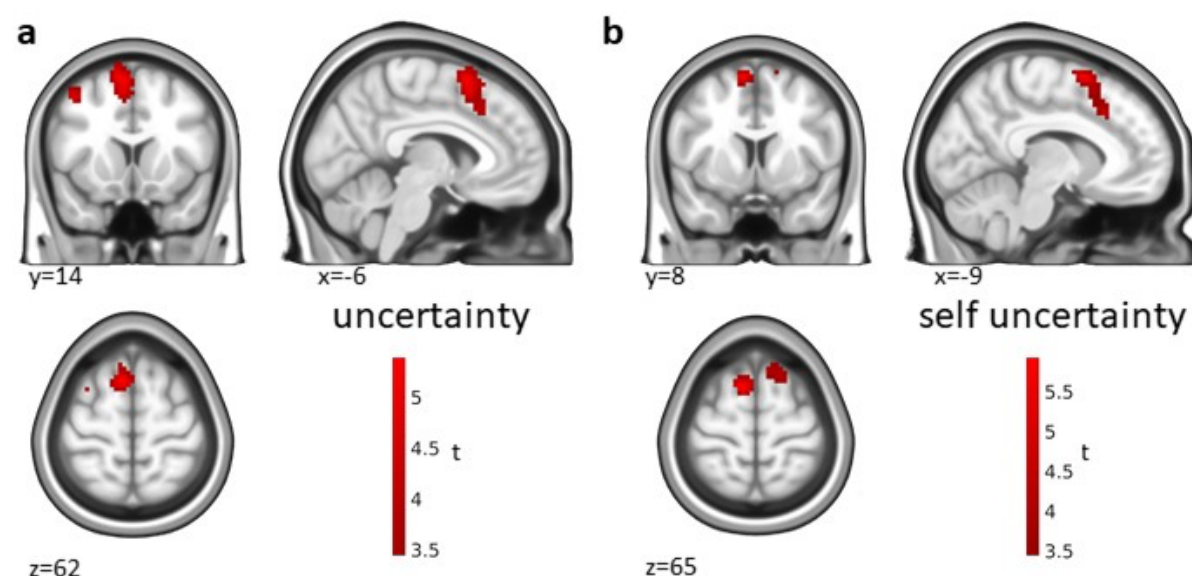

Figure 2. Self-referential processing. Activation for the contrast 'self > other' in the statement phase. Significance threshold p=0.05 cluster-level FWE-corrected with a cluster-defining threshold (CDT) of p=0.001 unc.
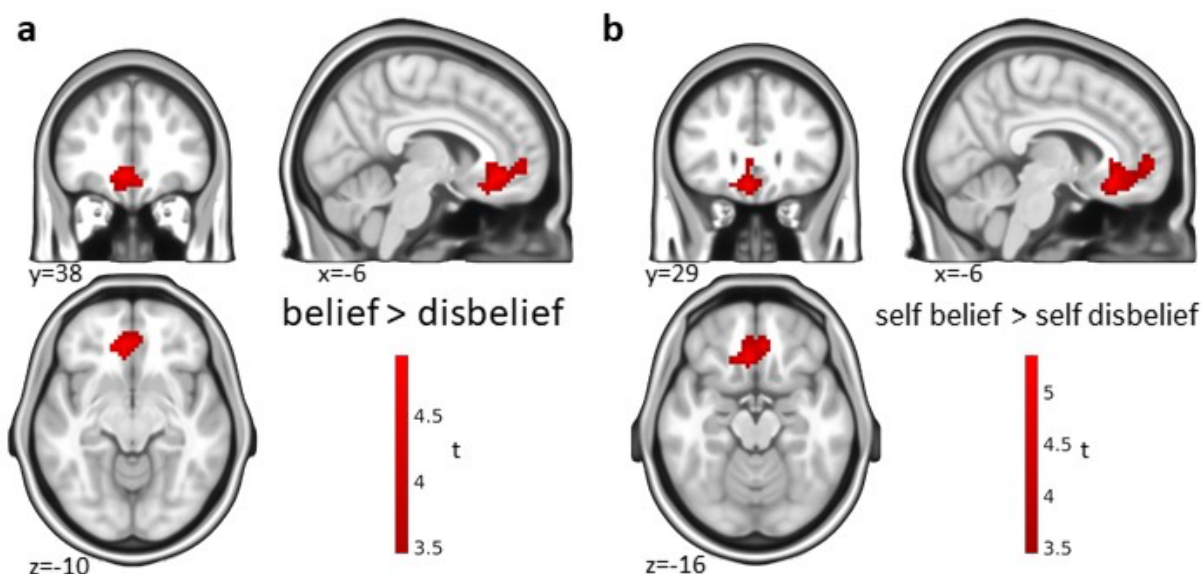
Figure 3. Belief processing. Activation for the contrast 'belief > disbelief' in the statement phase a) over all conditions and b) in the self-referential condition. Significance threshold p=0.05 cluster-level FWE-corrected with a cluster-defining threshold (CDT) of p=0.001 unc.

Figure 4. Uncertainty processing. Linear relationship (parametric modulation) of activation in the statement phase with uncertainty a) over all conditions and b) in the self-referential condition. Significance threshold p=0.05 cluster-level FWE-corrected with a cluster-defining threshold (CDT) of p=0.001 unc.
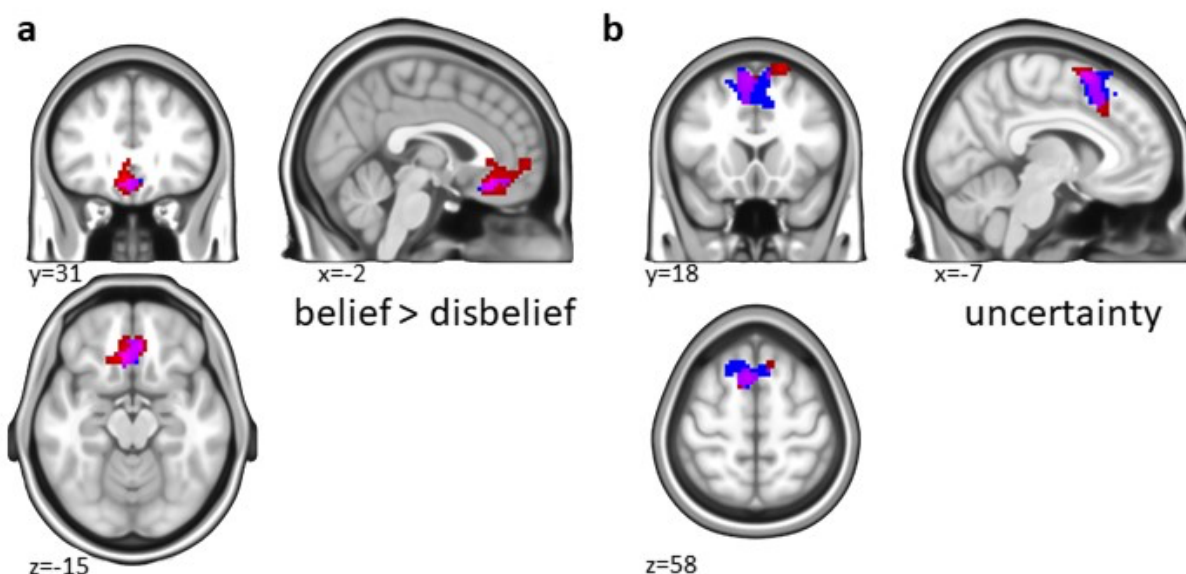
Figure 5. Overlap of self-referential belief processing with general belief processing in Gerchen et al., 2023. a) For the contrast 'belief > disbelief'. b) For the association with uncertainty. Red: Self-referential belief processing; Blue: General belief processing for statements about facts, politics, religion, conspiracy theories, and superstition reported in Gerchen et al., 2023. Magenta: Overlap of the effects

Tables

cluster-level		peak-level		
p_FWE	k_E	p_FWE	t	MNI (mm; x, y, z)
self > other				
<0.001	3523	<0.001	10	-6, 44, 11
		<0.001	8.30	-6, 32, -4
		<0.001	8.28	-9, 5, -7
<0.001	643	0.03	6.23	42, -55, -43
		0.065	5.85	24, -73, -22
		0.067	5.83	45, -64, -34
<0.001	305	0.25	5.14	-27, -70, -25
		0.29	5.05	-45, -67, -31
		0.73	4.38	-45, -58, -43

Table 1. Self-referential processing. fMRI results for the contrast 'self > other' in the statement phase. Significance threshold p=0.05 cluster-level FWE-corrected with a cluster-defining threshold (CDT) of p=0.001 unc.

cluster-level		peak-level		
p_{FWE}	k_E	p_{FWE}	t	MNI (mm; x, y, z)
belief > disbelief				
<0.001	224	0.44	4.96	-6, 38, -10
		0.59	4.75	-12, 32, -16
		0.69	4.62	-9, 53, -4
self belief > self disbelief				
0.096	60	0.21	5.36	-18, -10, -19
		0.94	4.12	-12, -1, -13
<0.001	269	0.21	5.36	-6, 29, -16
		0.32	5.13	6, 38, -16
		0.38	5.02	-9, 53, -1
0.096	60	0.56	4.74	9, 2, -13
		0.92	4.16	15, 8, -16
		0.99	3.62	24, 8, -19

Table 2. Belief processing. fMRI results for the contrast 'belief > disbelief' over all conditions and the contrast 'self belief > self disbelief' in the statement phase. In this case an explorative significance threshold of p=0.10 cluster-level FWE-corrected with a cluster-defining threshold (CDT) of p=0.001 unc. was used to include the small subcortical clusters in the table. Explorative results with 0.05 > p > 0.10 are shown in gray.

cluster-level		peak-level		
p_{FWE}	k_E	p_{FWE}	t	MNI (mm; x, y, z)
uncertainty				
<0.001	184	0.25	5.39	-6, 14, 62
		0.89	4.35	-9, 23, 38
		0.99	3.76	9, 17, 50
0.036	67	0.52	4.92	-39, 8, 50
self uncertainty				
<0.001	155	0.095	5.93	-9, 8, 65
		0.75	4.63	-9, 23, 38
		0.99	3.92	9, 23, 38
0.019	76	0.23	5.46	-39, 35, -1
		0.52	4.94	-51, 41, -10
		0.99	4.01	-48, 50, -10
0.035	65	0.34	5.23	18, 17, 68
		0.96	4.21	9, 20, 62

Table 3. Uncertainty processing. Linear relationship (parametric modulation) of activation in the statement phase with uncertainty over all conditions ('uncertainty') and in the self-referential condition ('self uncertainty'). Significance threshold p=0.05 cluster-level FWE-corrected with a cluster-defining threshold (CDT) of p=0.001 unc.

Author contributions
EB, IS, PK, & MFG planned and designed the study. EB, IS, & MFG created the study material. EB, IS, & MFG conducted the study and analyzed the data. EB, GK, PK, & MFG wrote the manuscript.

Conflicts of Interest
The authors declare no conflict of interest.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of personality and social psychology, 49*(6), 1621-1630. https://doi.org/10.1037/0022-3514.49.6.1621

Asp, E., Manzel, K., Koestner, B., Cole, C. A., Denburg, N. L., & Tranel, D. (2012). A neuropsychological test of belief and doubt: damage to ventromedial prefrontal cortex increases credulity for misleading advertising. *Frontiers in neuroscience, 6*, Article 100. https://doi.org/10.3389/fnins.2012.00100

Asp, E. W., Khan, L., Jonason, A., Adkins-Hempel, M., Warner, K., Pardilla-Delgado, E., Payne, J., Mullineaux, P., Denburg, N., & Tranel, D. (2022). Spinoza's make believe: propositions are regarded as true during comprehension. Available at SSRN 4086284 or http://dx.doi.org/10.2139/ssrn.4086284

Bain, A. (1859). *The emotions and the will*. John W. Parker and Son, West Strand. https://books.google.de/books?id=6DQCAAAAQAAJ

Barbey, A. K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex, 49*(5), 1195-1205. doi: 10.1016/j.cortex.2012.05.022

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain, 123*(11), 2189-2202. https://doi.org/10.1093/brain/123.11.2189

Bonner, M. F., & Price, A. R. (2013). Where Is the anterior temporal lobe and what does it do? *The Journal of Neuroscience, 33*(10), 4213-4215. https://doi.org/10.1523/JNEUROSCI.0041-13.2013

Bowen, R., Dmitriev, D., & Galperti, S. (2023). Learning from shared news: when abundant information leads to belief polarization. *The Quarterly Journal of Economics, 138*(2), 955–1000. https://doi.org/10.1093/qje/qjac045

Brown, M., Robinson, L., Campione, G. C., Wuensch, K., Hildebrandt, T., & Micali, N. (2017). Intolerance of uncertainty in eating disorders: a systematic review and meta-analysis. *European Eating Disorders Review, 25*(5), 329-343. https://doi.org/10.1002/erv.2523

Cabeza, R., & St Jacques, P. (2007). Functional neuroimaging of autobiographical memory. *Trends in Cognitive Sciences, 11*(5), 219-227. https://doi.org/10.1016/j.tics.2007.02.005

Chen, L., Ralph, M. A. L., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature human behaviour, 1*(3), 1-10. doi: 10.1038/s41562-016-0039

Cole, D. A., Howard, G. S., & Maxwell, S. E. (1981). Effects of mono- versus multiple-operationalization in construct validation efforts. *Journal of Consulting and Clinical Psychology, 49*(3), 395-405. https://doi.org/10.1037/0022-006X.49.3.395

Cooper, M. J., & Cowen, P. (2009). Negative self-beliefs in relation to eating disorder and depressive symptoms: different themes are characteristic of the two sets of symptoms in those with eating disorders and/or depression. *Journal of Cognitive Psychotherapy, 23*(2), 147-159. DOI: 10.1891/0889-8391.23.2.147

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351*, 1413-1420. https://doi.org/10.1098/rstb.1996.0125

Douglas, P. K., Lau, E., Anderson, A., Head, A., Kerr, W., Wollner, M., Moyer, D., Li, W., Durnhofer, M., & Bramen, J. (2013). Single trial decoding of belief decision making from EEG and fMRI data using independent components features. *Frontiers in Human Neuroscience, 7*, Article 392. https://doi.org/10.3389/fnhum.2013.00392

Dozois, D. J., & Rnic, K. (2015). Core beliefs and self-schematic structure in depression. *Current Opinion in Psychology, 4*, 98-103. https://doi.org/10.1016/j.copsyc.2014.12.008

Dumas, J. E., Johnson, M., & Lynch, A. M. (2002). Likableness, familiarity, and frequency of 844 person-descriptive words. *Personality and individual differences, 32*(3), 523-531. https://doi.org/10.1016/S0191-8869(01)00054-X

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10*(1), 48-58. https://doi.org/10.1038/nrn2536

Gao, T., Han, X., Bang, D., & Han, S. (2022). Cultural differences in neurocognitive mechanisms underlying believing. *Neuroimage, 250*, 118954. https://doi.org/10.1016/j.neuroimage.2022.118954

Gentes, E. L., & Ruscio, A. M. (2011). A meta-analysis of the relation of intolerance of uncertainty to symptoms of generalized anxiety disorder, major depressive disorder, and obsessive–compulsive disorder. *Clinical Psychology Review, 31*(6), 923-933. /https://doi.org/10.1016/j.cpr.2011.05.001

Gerchen, M. F., Glock, C., Weiss, F., & Kirsch, P. (2023). The truth is in there: belief processes in the human brain.

Gilbert, D. T. (1991). How mental systems believe. *American psychologist, 46*(2), 107-119. https://doi.org/10.1037/0003-066X.46.2.107

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of personality and social psychology, 65*(2), 221-233. https://doi.org/10.1037/0022-3514.65.2.221

Goel, V., & Dolan, R. J. (2003). Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning. *Neuroimage, 20*(4), 2314-2321. https://doi.org/10.1016/j.neuroimage.2003.07.027

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences, 15*(2), 56-67. https://doi.org/10.1016/j.tics.2010.12.004

Haghtalab, N., Jackson, M. O., & Procaccia, A. D. (2021). Belief polarization in a complex world: a learning theory perspective. *Proceedings of the National Academy of Sciences, 118*(19). https://doi.org/10.1073/pnas.2010144118

Han, X., Zhang, T., Wang, S., & Han, S. (2017). Neural correlates of believing. *Neuroimage, 156*, 155-165. https://doi.org/10.1016/j.neuroimage.2017.05.035

Harris, S., Kaplan, J. T., Curiel, A., Bookheimer, S. Y., Iacoboni, M., & Cohen, M. S. (2009). The neural correlates of religious and nonreligious belief. *PLoS one, 4*(10), Article e7272. https://doi.org/10.1371/journal.pone.0007272

Harris, S., Sheth, S. A., & Cohen, M. S. (2008). Functional neuroimaging of belief, disbelief, and uncertainty. *Annals of Neurology, 63*(2), 141-147. https://doi.org/10.1002/ana.21301

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe it or not: on the possibility of suspending belief. *Psychological Science, 16*(7), 566-571. https://doi.org/10.1111/j.0956-7976.2005.01576.x

Hornak, J., O'doherty, J., Bramham, J., Rolls, E., Morris, R., Bullock, P., & Polkey, C. (2004). Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience, 16*(3), 463-478. https://doi.org/10.1162/089892904322926791

Howlett, J. R., & Paulus, M. P. (2015). The neural basis of testable and non-testable beliefs. *PLoS one, 10*(5), Article e0124596. https://doi.org/10.1371/journal.pone.0124596

James, W. (1890/1910). *The Principles of Psychology*. Vol. II. New York: Henry Holt and Company. https://books.google.de/books?id=4GPSD2Oc5_YC

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports, 6*(1), Article 39589. https://doi.org/10.1038/srep39589

Kasper, L., Bollmann, S., Diaconescu, A. O., Hutton, C., Heinzle, J., Iglesias, S., Hauser, T. U., Sebold, M., Manjaly, Z.-M., & Pruessmann, K. P. (2017). The PhysIO toolbox for modeling physiological noise in fMRI data. *Journal of Neuroscience Methods, 276*, 56-72. https://doi.org/10.1016/j.jneumeth.2016.10.019

Kim, H. (2012). A dual-subsystem model of the brain's default network: self-referential processing, memory retrieval processes, and autobiographical memory retrieval. *Neuroimage, 61*(4), 966-977. https://doi.org/10.1016/j.neuroimage.2012.03.025

King, L. A., King, D. W., & Klockars, A. J. (1983). Dichotomous and multipoint scales using bipolar adjectives. *Applied Psychological Measurement, 7*(2), 173-180. https://doi.org/10.1177/014662168300700205

Koenigs, M., Barbey, A. K., Postle, B. R., & Grafman, J. (2009). Superior parietal cortex is critical for the manipulation of information in working memory. *Journal of Neuroscience, 29*(47), 14980-14986. https://doi.org/10.1523/JNEUROSCI.3706-09.2009

Kube, T., & Rozenkrantz, L. (2021). When beliefs face reality: an integrative review of belief updating in mental health and illness. *Perspectives on Psychological Science, 16*(2), 247-274. https://doi.org/10.1177/1745691620931496

Kuiper, N. A., & Rogers, T. B. (1979). Encoding of personal information: self–other differences. *Journal of personality and social psychology, 37*(4), 499-514. https://doi.org/10.1037/0022-3514.37.4.499

Kurczek, J., Wechsler, E., Ahuja, S., Jensen, U., Cohen, N. J., Tranel, D., & Duff, M. (2015). Differential contributions of hippocampus and medial prefrontal cortex to self-projection and self-referential processing. *Neuropsychologia, 73*, 116-126. https://doi.org/10.1016/j.neuropsychologia.2015.05.002

Lazzari, R., Fioravanti, M., & Gough, H. G. (1978). A new scale for the adjective check list based on self vs. ideal-self discrepancies. *Journal of Clinical Psychology, 34*(2), 361-365. https://doi.org/10.1002/

1097-4679(197804)34:2<361::AID-JCLP2270340218>3.0.CO;2-6

Lemogne, C., Gorwood, P., Bergouignan, L., Pélissolo, A., Lehéricy, S., & Fossati, P. (2011). Negative affectivity, self-referential processing and the cortical midline structures. *Social Cognitive and Affective Neuroscience, 6*(4), 426-433. https://doi.org/10.1093/scan/nsq049

Marques, J. F., Canessa, N., & Cappa, S. (2009). Neural differences in the processing of true and false sentences: insights into the nature of 'truth' in language comprehension. *Cortex, 45*(6), 759-768. https://doi.org/10.1016/j.cortex.2008.07.004

Matsumoto, K., & Tanaka, K. (2004). The role of the medial prefrontal cortex in achieving goals. *Current Opinion in Neurobiology, 14*(2), 178-185. https://doi.org/10.1016/j.conb.2004.03.005

NeuroImaging & Surgical Technologies Lab. (2021). MNI ICBM 152 non-linear 6th-generation symmetric average brain stereotaxic registration model. http://nist.mni.mcgill.ca/?p=858

Nguyen, D. T., Wright, E. P., Dedding, C., Pham, T. T., & Bunders, J. (2019). Low self-esteem and its association with anxiety, depression, and suicidal ideation in Vietnamese secondary school students: a cross-sectional study. *Frontiers in Psychiatry, 10*, Article 698. https://doi.org/10.3389/fpsyt.2019.00698

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain - a meta-analysis of imaging studies on the self. *Neuroimage, 31*(1), 440-457. https://doi.org/https://doi.org/10.1016/j.neuroimage.2005.12.002

Orth, U., & Robins, R. (2013, 12/03). Understanding the link between low self-esteem and depression. *Current Directions in Psychological Science, 22*(6), 455-460. https://doi.org/10.1177/0963721413492763

Papenberg, M., & Klau, G. W. (2021). Using anticlustering to partition data sets into equivalent parts. *Psychological Method*s, *26*(2), 161–174. https://doi.org/10.1037/met0000301

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*(12), 976-987. https://doi.org/10.1038/nrn2277

Pott, J., & Schilbach, L. (2022). Tracking and changing beliefs during social interaction: where computational psychiatry meets cognitive behavioral therapy. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.1010012

Schonbach, P. (1972). Likableness ratings of 100 German personality-trait words corresponding to a subset of Anderson's 555 trait words. *European Journal of Social Psychology, 2*(3), 327–333. https://doi.org/10.1002/ejsp.2420020309

Seitz, R. J. (2022). Believing and beliefs - neurophysiological underpinnings. *Frontiers in Behavioral Neuroscience, 16*. https://doi.org/10.3389/fnbeh.2022.880504

Seitz, R. J., & Angel, H.-F. (2012). Processes of believing - a review and conceptual account. *Reviews in the Neurosciences, 23*(3), 303-309. https://doi.org/10.1515/revneuro-2012-0034

Seitz, R. J., Paloutzian, R. F., & Angel, H.-F. (2018). From believing to belief: a general theoretical model. *Journal of Cognitive Neuroscience, 30*(9), 1254-1264. https://doi.org/10.1162/jocn_a_01292

Simmons, W. K., & Martin, A. (2009). The anterior temporal lobes and the functional architecture of semantic memory. *Journal of the International Neuropsychological Society, 15*(5), 645-649. https://doi.org/10.1017/S1355617709990348

Smith, B., Fowler, D. G., Freeman, D., Bebbington, P., Bashforth, H., Garety, P., Dunn, G., & Kuipers, E. (2006). Emotion and psychosis: links between depression, self-esteem, negative schematic beliefs and delusions and hallucinations. *Schizophrenia research, 86*(1-3), 181-188. https://doi.org/10.1016/j.schres.2006.06.018

Sui, J., & Humphreys, G. W. (2015). The integrative self: how self-reference integrates perception and memory. *Trends in Cognitive Sciences, 19*(12), 719-728. /https://doi.org/10.1016/j.tics.2015.08.015

The FIL Methods Group. (2014). SPM12 manual. SPM12. https://www.fil.ion.ucl.ac.uk/spm/doc/spm12_manual.pdf

Tsapkini, K., Frangakis, C. E., & Hillis, A. E. (2011). The function of the left anterior temporal pole: evidence from acute stroke and infarct volume. *Brain, 134*(10), 3094-3105. https://doi.org/10.1093/brain/awr050

Vandenberghe, R., Nobre, A. C., & Price, C. J. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience, 14*(4), 550-560. https://doi.org/10.1162/08989290260045800

Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013, Aug). Validating the truth of propositions: behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience, 8*(6), 647-653. https://doi.org/10.1093/scan/nss042