Evidence for Spinozan "Unbelieving" in the Right Inferior Prefrontal Cortex

Regan M. Bernhard[1], Steven M. Frankland[2], Dillon Plunkett[3,4], Beau Sievers[4,5], and Joshua D. Greene[4]

Abstract

■ Humans can think about possible states of the world without believing in them, an important capacity for high-level cognition. Here, we use fMRI and a novel "shell game" task to test two competing theories about the nature of belief and its neural basis. According to the Cartesian theory, information is first understood, then assessed for veracity, and ultimately encoded as either believed or not believed. According to the Spinozan theory, comprehension entails belief by default, such that understanding without believing requires an additional process of "unbelieving." Participants (n = 70) were experimentally induced to have beliefs, desires, or mere thoughts about hidden states of the shell game (e.g., believing that the dog is hidden in the upper right corner). That is, participants were induced to have specific "propositional attitudes" toward specific "propositions" in a controlled way. Consistent with the Spinozan theory, we found that thinking about a proposition without believing it is associated with increased activation of the right inferior frontal gyrus. This was true whether the hidden state was desired by the participant (because of reward) or merely thought about. These findings are consistent with a version of the Spinozan theory whereby unbelieving is an inhibitory control process. We consider potential implications of these results for the phenomena of delusional belief and wishful thinking. ■

INTRODUCTION

In 2015, ecologists working off the coast of Bali discovered a species of aquatic snake that uses electricity to stun its prey, much as electric eels do. The previous sentence is false. There are no such snakes. However, for a moment, you likely believed in them. What changed in your brain when you went from believing in electric snakes to merely thinking about them?

As this example illustrates, humans can represent an idea about the world ("There are electric snakes near Bali") as something that is true, false, possibly true, preferably true, or, in this case, merely as an idea. These distinct ways of relating to ideas are known as "propositional attitudes" (McKay & Nelson, 2014). Our focus here is on the distinction between representing an idea, or "proposition," as true (believing), and representing an idea without believing that it is true. It is unknown whether the ability to entertain ideas without believing in them is an ability we share with other animals. In humans, however, this ability is likely essential for high-level cognitive functions such as planning under conditions of uncertainty and hypothesis testing, whether in science or everyday life. Furthermore, the inability to prevent one's mere thoughts from becoming beliefs may underlie mental disorders involving delusion, such as schizophrenia (Bronstein, Pennycook, Bear, Rand, & Cannon, 2019; Bear, Fortgang, Bronstein, & Cannon, 2017; Dudley, Taylor, Wickham, & Hutton, 2016; Evans, Averbeck, & Furl, 2015; Moritz & Woodward, 2005).

Beliefs are central to human cognition, but the mechanisms that distinguish belief from mere representation have received surprisingly little attention. It is worth pausing, then, to distinguish the present project from relevant research that is likely more familiar. There has been extensive research on how humans represent the beliefs of others ("mentalizing" or "theory of mind"; e.g., Decety & Lamm, 2007; Saxe & Baron-Cohen, 2006; Saxe & Kanwisher, 2003) as well as how people represent their own beliefs (a form of "metacognition"; e.g., Mazzoni & Nelson, 1998; Metcalfe & Shimamura, 1994). There is also considerable research on the representation of probability and uncertainty about states of the world (e.g., Delgado, Miller, Inati, & Phelps, 2005; Glimcher, 2003; Critchley, Mathias, & Dolan, 2001), which can be understood as the encoding of degrees of belief. Finally, as indicated above, perception typically results in belief, and memory enables the retention of beliefs. As the foregoing suggests, beliefs are cognitively ubiquitous, so much so that one might question whether belief per se can be a meaningful cognitive topic. In our view, it is not so much belief that is a distinct and underappreciated topic, but rather the ability to represent ideas without believing them.

Here, we study this ability using functional neuroimaging and a novel behavioral task, inspired by the street-side

[1]Boston College, [2]Princeton University, [3]Northeastern University, [4]Harvard University, [5]Stanford University

https://doi.org/10.1162/jocn_a_01964

"shell game." Our shell game involves objects hidden in various locations, and participants can be systematically induced to have beliefs, desires, or mere thoughts about where the hidden objects are. Critically, this gives us experimental control over participants' propositional attitudes, and that, in turn, enables us to dissociate propositional attitudes from propositional content. This degree of experimental control distinguishes the shell game from prior paradigms used to study the neural mechanisms of belief and/or desire. This task also avoids confounds related to metacognition, perception, and memory. We use this method to test two cognitive models of belief, each with a long history (Gilbert, Tafarodi, & Malone, 1993; Gilbert, 1991).

According to the seventeenth century philosopher René Descartes, ideas are first comprehended and then assessed for truth, such that ideas deemed true are subsequently encoded as beliefs (Mandelbaum, 2014; Gilbert, 1991; Descartes, 1984). Descartes' contemporary, Baruch Spinoza, proposed an alternative theory. Spinoza argued that comprehension entails belief by default, implying that understanding without believing requires an additional step—a process we call "unbelieving" (Mandelbaum, 2014; Gilbert, 1991; Spinoza, 1677/1982). This could involve tagging ideas as not (necessarily) true or it could require inhibiting belief processes. The Cartesian theory, by contrast, posits no distinctive extra step for non-belief. It requires only that our brains somehow encode the epistemic status of representations as beliefs or non-beliefs, with varying degrees of certainty. We now discuss the evidence for each of these theories.

Spinozan Belief

Evidence for the Spinozan view of belief comes primarily from the phenomenon of truth bias—the default tendency to encode information as true (Vrij, 2008; Bond & DePaulo, 2006; Levine, Park, & McCornack, 1999). People are influenced by false information even when the information is explicitly presented as false (Thorson, 2015; Guenther & Alicke, 2008; Schul & Burnstein, 1985; Anderson, 1983; Anderson, Lepper, & Ross, 1980). People also remember false information as true more often than the reverse (Peter & Koch, 2016). Critically, truth bias appears to be magnified under cognitive load and time pressure (Gilbert et al., 1993; Gilbert, Krull, & Malone, 1990), suggesting that non-belief rather than belief requires additional cognitive effort. Further support for the Spinozan view comes from the illusory truth effect, whereby prior exposure to an idea makes it seem more true (Dechêne, Stahl, Hansen, & Wänke, 2010; Begg, Anas, & Farinacci, 1992). The illusory truth effect is diminished by critical engagement and increased by more superficial processing (Hawkins & Hoch, 1992). These findings support Spinoza's speculation that belief is an automatic consequence of comprehension and that understanding something without believing it requires an additional, cognitively demanding process of "unbelieving."

The behavioral evidence described above suggests that entertaining ideas without believing them may require cognitive control and therefore may recruit control-related brain regions such as the dorsolateral pFC, frontopolar cortex, inferior frontal gyrus (IFG), and anterior cingulate cortex (Niendam et al., 2012; Aron, Behrens, Smith, Frank, & Poldrack, 2007; Cole & Schneider, 2007; Chein & Schneider, 2005). Consistent with this, participants tasked with evaluating the veracity of statements exhibit increased activity in the frontopolor cortex when considering false (vs. true) statements (Marques, Canessa, & Cappa, 2009). Furthermore, evaluating statements as false is associated with reduced activation in the ventral medial prefrontal cortex (vmPFC; Harris et al., 2009; Harris, Sheth, & Cohen, 2008). The vmPFC is part of the default mode network, which typically exhibits reduced activity when people engage in cognitively demanding tasks (Gusnard, Akbudak, Shulman, & Raichle, 2001; Mazoyer et al., 2001). This reduction in vmPFC activity for rejected statements could, then, potentially result from the inhibition of a belief-related process. Likewise, individuals with damage to the vmPFC are more likely than control participants (both healthy controls and those with damage outside the frontal cortex) to believe misleading advertisements, even when the advertisements include a disclaimer about their falsehood (Asp et al., 2012). Again, this is consistent with the idea that the vmPFC is implicated in controlling or inhibiting belief.

In the neuroimaging and patient studies described above, participants explicitly reported on what they did or did not believe. Thus, it is unclear whether the observed neural responses reflect the participants' beliefs (or non-belief) or whether these responses instead reflect the metacognitive processes involved in reporting on beliefs. Avoiding this problem, Goel and Dolan (2003) examined the neural mechanisms of belief without having participants explicitly reflect or report on their beliefs. Participants evaluated the quality of arguments in which the conclusions were either consistent with the participants' beliefs or not. The authors found increased activation in the right IFG when participants correctly accepted valid arguments with conclusions that violated their beliefs (e.g., "No harmful substances are natural; All poisons are natural; [therefore] No poisons are harmful"), than when they failed to accept such arguments. More recent follow-up research found that repetitive transcranial magnetic stimulation to the right IFG increased the likelihood that participants accepted invalid arguments when the components conformed to their beliefs (Tsujii, Sakatani, Masuda, Akiyama, & Watanabe, 2011; Tsujii, Masuda, Akiyama, & Watanabe, 2010). In each of these studies, however, belief was confounded with propositional content. In other words, whether participants believed what they read depended on the plausibility of what they were reading. Although this is to be expected in everyday life, it poses a challenge for studies aiming to understand the neural mechanisms of belief. Such designs leave open the possibility that neural activity associated with belief reflects the contents of relatively believable propositions rather than the mental process of believing.

The Spinozan view raises parallel questions about a different propositional attitude, namely, *desire*. One can (and often does) desire that something be true without believing that it is true. Nevertheless, a strong form of the Spinozan theory predicts that desiring that something be true will, by default, generate a belief that it *is* true, such that this belief will persist if it is not overridden. In other words, the Spinozan theory, in addition to explaining our susceptibility to phenomena such as the illusory truth effect (which applies to neutral content), may also explain our susceptibility to *wishful thinking* (Mandelbaum, 2014; Windschitl, Scherer, Smith, & Rose, 2013; Aue, Nusbaum, & Cacioppo, 2012; Krizan & Windschitl, 2009; Babad, 1997). Consistent with this, there is some evidence that wishful thinking occurs automatically (Cahill, 2015).

Cartesian Belief

In contrast to the Spinozan view of belief as the representational default, the Cartesian view posits that information is first understood, then subsequently assessed for veracity, and ultimately encoded as either believed or not believed (with varying degrees of confidence; Gilbert, 1991; Descartes, 1984). Some advocates of the Cartesian view argue that the effect of cognitive load on truth bias can be explained as an effect of increased uncertainty, rather than the automaticity of belief (Street & Kingstone, 2017). They find that the effect of time pressure on the tendency to report false statements as true disappears when participants can respond that they are uncertain of the statement's truth value (Street & Kingstone, 2017; Street & Richardson, 2015). If belief is automatic, then cognitive load or time pressure should increase reports of belief, regardless of whether the alternative response is "false" or "not sure." By contrast, if what is automatically engaged is a Cartesian truth-assessment process, then time pressure or increased load should increase reports of uncertainty (Street & Kingstone, 2017).

Other research raises doubts about the Spinozan theory by questioning the necessity of cognitive load or time pressure to evoke truth bias (Fiedler, Armbruster, Nickel, Walther, & Asbeck, 1996; Fiedler, Walther, Armbruster, Fay, & Naumann, 1996). Participants more often mistake false statements as true, rather than the reverse, even in the absence of cognitive load and even when the statements are explicitly tagged as false (Pantazi, Kissine, & Klein, 2018). According to these researchers, this implies that truth bias is not the product of automatic belief (Pantazi et al., 2018).

Finally, some researchers claim that variability in truth bias counts against the idea that belief is the immutable default of comprehension, countering the provocative Spinozan claim that "You can't not believe everything you read" (Gilbert et al., 1993). Truth bias appears to be moderated by context (Dechêne et al., 2010), source credibility (Nadarevic & Erdfelder, 2013; Henkel & Mattson, 2011), the degree to which the statements are informative (Hasson, Simmons, & Todorov, 2005; Fillenbaum, 1966), and background knowledge or beliefs (Richter, Schroeder, & Wöhrmann, 2009; but see Fazio, Brashier, Payne, & Marsh, 2015). Consistent with the Cartesian view, these findings suggest that believing is not always the default. Rather, Cartesian belief theorists argue that there is an efficient evaluation process that sets the default based on background knowledge, context, and so forth.

If truth bias is not the result of automatic belief, what explains it? The most common, non-Spinozan explanation for truth bias appeals to processing fluency, that is, ease of comprehension. Under this view, repetition makes statements easier to process, which leads individuals to conclude that such statements are more likely to be true, even when they are known to be false (Unkelbach & Stahl, 2009; Unkelbach, 2007). Consistent with this view, some studies have reported illusory truth effects as a result of fluency, even in the absence of prior exposure: Statements written in easy-to-read fonts are more often rated as true (Unkelbach, 2007; Reber & Schwarz, 1999).

Confidence, Probability, and Reward Anticipation

Considerable prior research has investigated the neural bases of confidence (e.g., Kiani & Shadlen, 2009; Kepecs, Uchida, Zariwala, & Mainen, 2008; Aron et al., 2007; Hampton & O'Doherty, 2007; Knutson, Rick, Wimmer, Prelec, & Loewenstein, 2007; Heekeren, Marrett, Bandettini, & Ungerleider, 2004; Glimcher, 2003; Shadlen & Newsome, 1996, 2001), uncertainty (Platt & Huettel, 2008; Glimcher, 2003), and representations of probabilities (see Ma & Jazayeri, 2014 and Meyniel, Sigman, & Mainen, 2015, for comprehensive reviews). However, this work has focused on identifying the neural architecture responsible for tracking and contrasting degrees of confidence and the representations of varying probabilities (e.g., high vs. low). Here, however, we are not contrasting high versus low degrees of belief, but rather contrasting belief as a mental state with different mental states. More specifically, we are asking how believing that the world is a certain way differs from (a) wanting the world to be that way and (b) merely thinking about the world as being that way.

Likewise, many studies have examined degrees of desire in different forms. Most often, desire has been conceptualized in the literature as predictive value tracking or stimulus valuation (e.g., Chib, Rangel, Shimojo, & O'Doherty, 2009; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Kable & Glimcher, 2007; Plassmann, O'Doherty, & Rangel, 2007; Tom, Fox, Trepel, & Poldrack, 2007; Sugrue, Corrado, & Newsome, 2005), as reward uncertainty, (e.g., O'Neill & Schultz, 2010; Yacubian et al., 2007; Abler, Walter, Erk, Kammerer, & Spitzer, 2006; Dreher, Kohn, & Berman, 2006; Fiorillo, Tobler, & Schultz, 2003; Critchley et al., 2001), or as reward anticipation (see Knutson & Greer, 2008, for a meta-analysis). However, this prior work has focused primarily on neural responses to variability in reward amount or likelihood. In the present research, we are not contrasting high levels of desire with low levels—or even no desire at all. Rather, we are contrasting desiring that the world be a certain way with believing that it is that way or merely thinking about it is being that way.

In short, many studies have examined the neural mechanisms associated with encoding degrees of belief or degrees of desire, but they have not contrasted these propositional attitudes with each other while controlling for the propositional content of the belief/desire. Nor have they contrasted belief or desire with mere thought, where the propositional content remains the same, but both belief and desire are absent. In the present work, we do just that, focusing on the differences between the neural instantiation of distinct propositional attitudes, rather than variation within the attitudes of belief or desire.

The Present Research

The present research examines the neural mechanisms of believing and withholding belief, testing predictions made by the Spinozan and Cartesian theories. We use fMRI coupled with our shell game task, which systematically induces participants to form beliefs, desires, or mere thoughts about the location of a hidden target object. As noted above, this strategy has several advantages over prior research. First, and most important, this strategy provides experimental control over what is believed, desired, or merely thought about. Once again, this enables us to dissociate propositional attitudes from propositional content. Second, our task does not require participants to report on their beliefs or desires, minimizing confounds related to metacognitive processing. Third, participants do not have direct perceptual experience of the propositions in question (i.e., seeing the target object in the target location). This distinguishes the present research from the vast body of research on perception and memory and allows us to focus specifically on the mechanisms of believing and withholding belief.

We directly test the Spinozan theory by searching for neural responses consistent with "unbelieving," the active withholding of belief. This inhibitory account contrasts with an alternative Spinozan account whereby unbelieving is executed by "tagging" beliefs as not true, whereas representing a belief as true requires no special tag. Although both inhibition and "tagging" could result in non-belief, they are different in much the way that censorship differs from applying a warning label. The former involves the prevention of a process, whereas the latter involves the appending of information. Evidence for the belief inhibition version of the Spinozan account would entail the engagement of inhibitory control processes dependent on the prefrontal cortex (Munakata et al., 2011; Dias, Robbins, & Roberts, 1997) when participants represent a proposition without believing that it is true. We also test a version of the Cartesian theory of belief by searching for activation that is specific to believing. (The Cartesian theory does not necessarily entail the activation of a specific *process* for belief, although the distinction between believed and non-believed propositions must be encoded somehow.)

As suggested above, a strong version of the Spinozan theory predicts that "unbelieving" processes will be engaged under two different conditions of non-belief: (1) when participants think about a proposition while neither believing that it is true nor desiring that it be true, and (2) when participants desire that a proposition be true without believing that it is true. If both predictions are confirmed, this would provide evidence that "wishful thinking" is a special case of truth bias, resulting from a failure of Spinozan belief inhibition.

METHODS

Participants

Eighty-seven people (47 women; ages 18–64 years) participated in this study. All were native English speakers, right-handed, and had no history of neurological conditions. Of these, 8 were excluded for poor performance (less than 70% accuracy) on the attention check task (see details below). Eight other participants were excluded for excessive motion during scanning, which was defined as 2 *SD*s from the group mean on two of the three following motion parameters: average absolute motion per run, average number of movements greater than 0.5 mm per run, and average per run slice-wise signal-to-noise ratio. Finally, one participant was excluded because of technical issues leading to incomplete data. The remaining 70 participants were included in all subsequent analyses described throughout the article.

Data were collected in two batches. After scanning the first 30 usable participants, the data were analyzed. We then resumed scanning until we had collected data from an additional 40 participants who met our inclusion criteria. This target sample size of 70 usable participants was determined after the collection and analysis of the data from the first 30 participants and was preregistered (Open Science Framework [https://osf.io/gs82p/]). The sample size (n = 30) for the original, exploratory sample was not based on a power analysis, as we had no reliable indications about likely effect sizes. However, the second sample size (n = 40) was based on a power analysis for a test of object classification accuracies in the left parahippocampal gyrus using multivariate pattern analysis. This sample was chosen to obtain 90% power to detect an effect (average classification accuracy across participants) of the size that we originally observed in our first sample.

Although our power analyses were not conducted with the combined 70-participant sample in mind, the larger sample, along with our use of simpler univariate methods, makes the analyses reported here better powered than the analyses targeted in our power analysis for the second sample. Our preregistered analysis plan included provisions for both a two-study exploratory/confirmatory analysis structure and an analysis of the full 70-participant data set. The former yielded suggestive findings, but key results did not survive stringent whole-brain correction within either subset. Nor did they survive an ROI-based analysis of the second subset. Here, we report on our preregistered analyses (with whole-brain correction) using the full set of 70 participants.

Stimuli

We developed a novel dynamic visual paradigm modeled after the classic street-side shell game (see Figure 1). The shell game task allows us to freely and arbitrarily manipulate the propositional attitudes (i.e., believing, desiring, or mere thinking) that participants have about different propositions (i.e., different positions of concealed objects). Participants view images of four objects (a dog, a mop, a snake, and a hose). After 5 sec, blue squares cover each object and a verbal cue indicates which object is the "target" on that trial, as well as which type of trial it is going to be ("track," "bonus," or "think"; described below). Then, the concealed objects are shuffled and distributed to each of the four corners of the screen.

Trials have three types: "track," "bonus," and "think," which induce belief, desire, and mere thinking, respectively. On track trials, participants track the location of the concealed target object. On these trials, the objects are shuffled in such a way that the (concealed) target object is easy to follow, and its final location is clear to the participant. Thus, during a track trial, participants have a *belief* about the target object's final location. However,

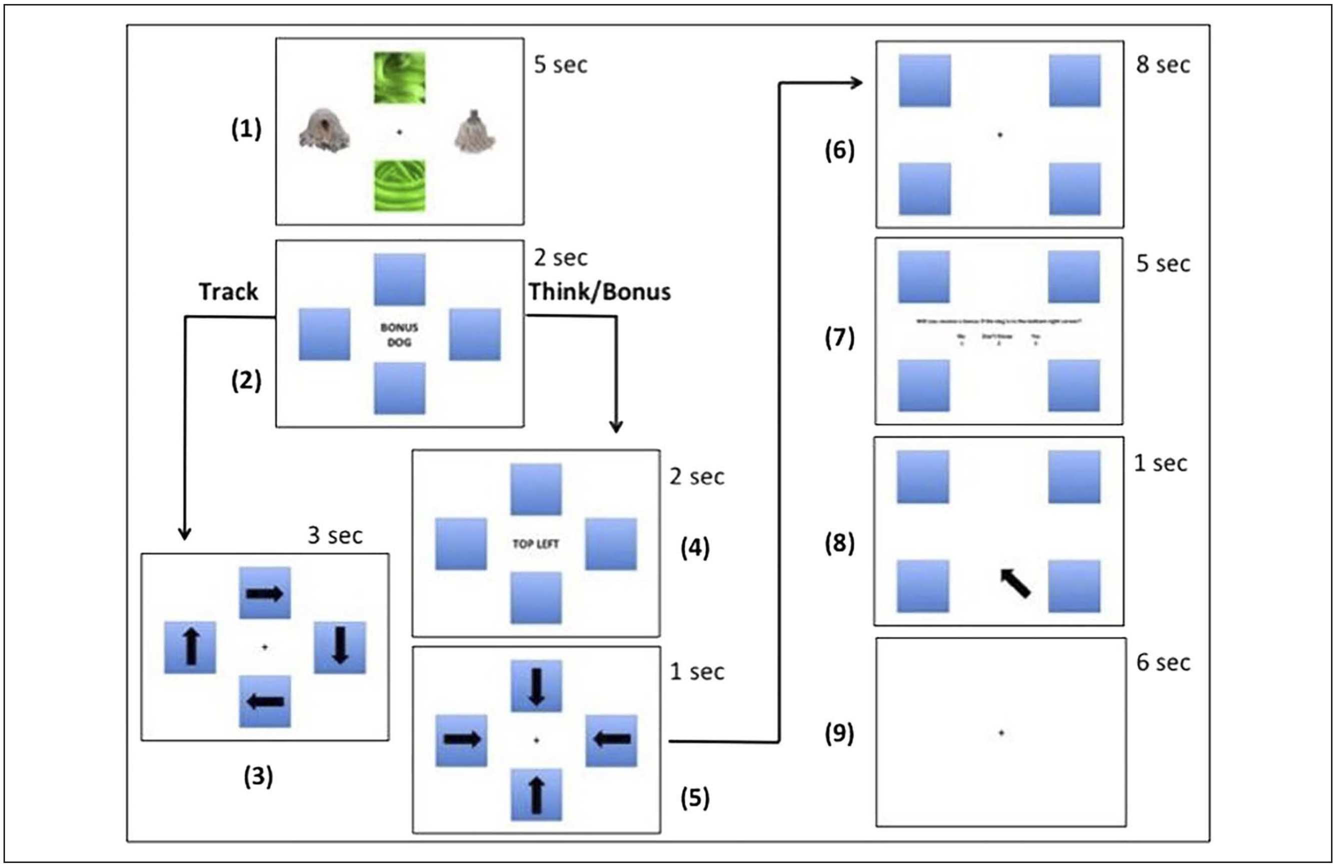


Figure 1. Task sequence and timing. (1) The objects are randomly placed in one of four locations (left, right, top, and bottom) for 5 sec. During this time, participants must learn the starting location of each object. (2) The objects are covered (for 2 sec), and participants learn the trial type (track, think, or bonus) and which of the four objects is the target object. (3) In track trials (belief condition), the blue squares are then shuffled. Over the course of 3 sec, the blue squares are either moved to their final locations or participants' view arrows pointing to the squares' final locations. (4) In the think trials (mere thought condition) and bonus trials (desire condition), a screen displays the target location for 2 sec. (5) The blue squares are then shuffled over the course of 1 sec either by having the squares move around the screen or by having arrows point to the squares' next location. Importantly, the objects in the think and desire/bonus trials are impossible to track, giving participants no knowledge of their final locations. (6) In all three trial types, the shuffle is followed by an 8-sec delay period. It is during this time period that all of our critical analyses are performed. (7) The delay period is followed by the attention check question, which participants have 5 sec to answer. (8) An arrow is then displayed on the screen for 1 sec, showing the final location of the target object. (9) Each trial ends with a 6-sec intertrial interval.

critically, they are given no special reason to desire that it be there. Accordingly, these trials constitute our "belief" condition. On "bonus" trials, participants are verbally informed about a target location in addition to a target object. Participants receive a $5 bonus if the target object ends up in the target location on a bonus trial. As they watch, the objects are shuffled in such a way that it is impossible to track the target object. Thus, participants desire that the target object be in the target location, but they are given no (good) reason to believe that the target object is in that location rather than any other. These trials constitute our "desire" condition. Finally, "think" trials closely resemble bonus trials in structure and visual content. After being informed about the trial type and the target object, participants are given a target location and are told to merely think about the target object as being in the target location. Critically, in the think condition, participants do not receive a bonus if the target object ends up in the target location, or anywhere else. Consequently, participants in the think condition have little reason to believe that the target object is in the target location; nor do they have reason to desire that the target object be in the target location.

To confirm that we successfully manipulated the degree to which participants had beliefs and desires in the relevant conditions, we asked our first 30 participants to complete a follow-up survey designed to evaluate the degree to which they believed and desired the target outcome in each condition. Specifically, to measure how strongly they believed that the target object was in the target location, we asked participants, "When you [had to track the target object to/had to think about the target object being in/got a bonus if the target object went to] the target location, how strongly did you believe it was there?" Participants responded on a 1–5 scale anchored at *not very much* and *very much*. Consistent with our expectation that participants held a stronger belief about the location of the target object in the track trials than in the bonus or think trials, we found that responses were significantly higher for the track trials ($M = 4.59, SD = 0.68$) than in the bonus ($M = 2.72$, $SD = 1.25$; $t(28) = 6.37$, $p < .0001$) or think trials ($M = 2.21, SD = .94$; $t(28) = 11.84, p < .0001$). However, in both desire and think trials, participants' average responses fell at the midpoint of the scale, suggesting that they were not operating with a complete absence of belief in these conditions. Nevertheless, these data indicate that the belief condition produces beliefs about the object's location far more than the desire and think conditions.

This survey also included a measure of how strongly participants desired that the target object be in the target location. Specifically, we asked the first set of participants (n = 30), "On a [Bonus/Think/Track] trial, how much would you have paid to have the target object end up in the target location?" Participants responded on a $0–2.00 scale in $0.50 increments (scale values ranging from 1–5). On average, participants were willing to pay $1.54 (out of $2.00) to have the desired outcome in bonus trials, but only $0.47 to have the desired outcome in think trials and $0.58 for the desired outcome in track trials. Paired-samples t tests confirmed that participants were willing to pay significantly more for the desired outcome in bonus than think trials, $t(28) = 10.61, p < .0001$, and in bonus than track trials, $t(28) = 6.98, p < .0001$. Moreover, one-sample t tests confirmed that participants' responses in think and track trials were not significantly different from $0 (think: $t(28) = 1.98, p = .057$; track: $t(28) = 1.94, p = .062$), but was for bonus trials, $t(28) = 12.37, p < .0001$. Given these results, we felt confident that participants were experiencing meaningful differences in their levels of belief and desire between conditions.

Because the shuffles in belief trials are always trackable, whereas shuffles in the desire and think trials are not, shuffle trackability is confounded with attitude type. The shuffles occur before the extended delay period on which our main analyses are based. To ensure that incidental perceptual differences in shuffle type during the prior period are not responsible for effects observed during the critical delay period, we generated two different shuffle styles. In one, the concealed objects visibly rotate around the screen. In the other, arrows appear on the screen indicating the final location of each of the concealed objects (in track/belief trials), or pointing to the center of the screen (in bonus/desire and think trials). The concealed objects then disappear from and reappear in each of the four corners of the screen. Creating two shuffle styles allows us to perform follow-up analyses assessing whether features of the shuffle are responsible for effects observed during the delay period. Our main analyses, however, were performed collapsing across shuffle styles. We note that differences observed between the think and desire conditions cannot be explained by differences in shuffle style because these two conditions both employ the same, nontrackable shuffle.

After the objects are shuffled, participants view a delay screen for 8 sec with the four blue squares in each corner and a fixation cross in the center. All analyses were performed using the modeled hemodynamic response corresponding to this delay period only. During this time, participants view the same screen (four blue squares, one in each corner) in all three conditions. Thus, during the delay period, the perceivable stimuli are identical, eliminating potential perceptual confounds.

The delay screen is followed by an attention-check, which ensures that participants have encoded the relevant information on each trial. Specifically, participants read one of the following questions:

Is the [target object] in [specific location]?
Will you get a bonus if the [target object] is in [specific location]?
Were you asked to think about the [target object] being in [specific location]?

Participants respond using a 1–3 scale (1 = *No*, 2 = *Don't Know*, 3 = *Yes*) by pressing the corresponding button on a button box held in their right hands. These questions are randomly assigned to each trial so that participants only receive the question specifically probing the mental state (believing, desiring, or thinking about) induced by the current trial type on approximately 1/3 of the trials. This was done to avoid confounding the condition with the expectation of a specific question. Furthermore, although we always ask about the target object, we ask about the target location on only 50% of the trials. On the remaining trials, we ask about a random, non-target location. In this way, participants are required on all trials to attend to and remember the trial type, the target object, and the target location. Responses were coded as correct depending on the question, condition, and location that was asked about as described above (see Table 1).

Finally, an arrow is flashed on the screen showing the final location of the target object. This arrow is randomly placed in one of five locations on the screen to eliminate any effect of anticipating the arrow's location. In belief/track trials, the arrow tells participants whether they correctly tracked the object. In desire/bonus trials, it tells participants whether they will receive a bonus on that trial. In the think trials, the arrow merely informs the participant of the target object's final location.

Importantly, participants never see any of the objects in their final locations. The objects begin each trial centered near each edge of the screen but are covered before they are moved to their final locations in the four corners. Critically, this feature of our paradigm enables us to examine neural activity associated with beliefs, desires, or thoughts about hidden states that have not been directly perceived and that therefore cannot be remembered.

Experimental Procedure

The task consists of 12 runs of 12 trials each. Each target object and condition (belief, desire, or think) was pseudorandomly assigned on each trial, such that each object was the target object 3 times per run and each condition occurred 4 times per run. The order of presentation was randomized across trials within each run. The target location for each trial was also pseudorandomly assigned such that the target location was randomly assigned on each trial, but target locations were equalized over the entire experiment (random without replacement). Participants received a bonus in only 25% of the desire trials (12 out of 48) to minimize their ability to anticipate the final location of the target object in these trials. These 12 trials were randomly distributed throughout the experiment. A single trial took 30 sec, with a minimum 30-sec break between runs. The duration of the entire experiment was approximately 80 min.

fMRI Acquisition and Preprocessing

Neuroimaging was performed using a Siemens Prisma 3.0 T scanner with a 32-channel head coil at the Harvard Brain Sciences Center in Cambridge, MA. A high-resolution structural scan was performed before functional data acquisition using a 3-D magnetization prepared rapid gradient echo sequence (repetition time [TR] = 2530 msec,

Table 1. Participants Were Asked One Attention Check Question at the End of Each Trial

Question	*Condition*	*Location Specified in Question*	*Responses Accepted as Correct*
Is the [target object] in [specific location]?	Belief	Target location	Yes
		Non-target location	No
	Desire or think	Target location	Don't know
		Non-target location	Don't know
Will you get a bonus if the [target object] is in [specific location]?	Desire	Target location	Yes
		Non-target location	No
	Belief or think	Target location	No
		Non-target location	No
Were you asked to think about the [target object] being in [specific location]?	Think	Target location	Yes
		Non-target location	No
	Belief or desire	Target location	Yes or no
		Non-target location	No

These questions were randomly assigned to the trial. Although the questions always asked about the target object, 50% of the questions asked about a non-target location. Correct answers depended on the combination of question, condition, and specified location.

echo time = 1.69 msec, flip angle = 7°, field of view = 256 mm, slice thickness = 1.0 mm, 176 slices). The EPI pulse sequence for functional scans used a 2000 msec TR with 190TRs per functional run (echo time = 35 msec, flip angle = 80°, field of view = 207 mm, slice thickness = 2.2 mm, 69 slices). Stimuli were presented using the Psychtoolbox package for MATLAB (The MathWorks).

Data preprocessing was performed using an adaptation of AFNI's *afni_proc.py* program (https://afni.nimh.nih.gov/pub/dist/doc/program_help/afni_proc.py.html). The first five TRs were removed from each run. After performing despiking and slice time correction, each participant's EPI images were spatially registered to the first volume of the second run using cubic polynomial interpolation. Data were smoothed for the univariate analyses with a Gaussian kernel at 6.6-mm FWHM (the equivalent of three voxels). A mask was created to remove any voxels with more than 12 TRs with no data and was used for all subsequent single-subject analyses.

Univariate Analyses

The hemodynamic response function deconvolution for the BOLD signal corresponding to the critical delay period was performed using AFNI's *3dDeconvolve* with a BLOCK model set for an 8-sec event duration (length of the delay period on each trial). An ordinary least squares regression was performed, with condition (i.e., belief, desire, or think) as the primary regressors, and motion parameters entered as regressors of no interest. General linear tests on the contrasts of interest (belief vs. desire, belief vs. think, desire vs. think) were included in this analysis. Finally, the resultant individual participant whole-brain beta-maps were warped to Talairach (TLRC) space via a diffeomorphic nonlinear transformation using Advanced Normalization Tools (ANTs) (Avants, Tustison, & Song, 2009).

Group analysis was performed using one-sample *t* tests (with AFNI's *3dttest++* function) to identify clusters of voxels in which the average betas from the individual-participant general linear tests were significantly different from zero. Using the *Clustsim* flag in *3dttest++*, we ran the Monte Carlo simulation to perform cluster-wise correction for multiple comparisons on the resultant *t* maps. As specified in our preregistration, a threshold for statistical significance was set at a voxel-wise $p < .001$, and a cluster-wise corrected threshold of $p < .05$. To identify brain regions preferentially activated by the belief condition, we ran separate *t* tests for the belief versus desire contrasts and belief versus think contrasts. We then looked for clusters of overlap between significant voxels from both tests. To identify those brain regions preferentially activated by the desire condition, we ran separate *t* tests for the desire versus belief contrasts and desire versus think contrasts. We then looked for clusters of overlap between significant voxels from both tests. Finally, to identify neural activation associated with "unbelieving" (thinking about a proposition without belief), we looked for clusters of overlap between significant voxels for the desire versus belief, and think versus belief contrasts.

Multivariate Analyses

Our multivariate analyses had two objectives. The first was to identify regions that encoded the identity of the target objects or target locations across the attitude condition. To this end, at the participant level, a single image for each trial was created by averaging over the temporal interval from 6 to 14 sec after the start of the critical 8-sec delay period. Using this data, we ran searchlight analyses (Kriegeskorte & Bandettini, 2007) implemented with the Searchmight Toolbox (Pereira & Botvinick, 2011). A cube with a 2-voxel (6.6 mm) radius was centered at each voxel, and MATLAB's built-in Gaussian naive Bayes classifier (implemented with the *NaiveBayes.fit* function) was trained in each searchlight neighborhood to classify either the target object (mop, dog, snake, or hose) or target location (top–left, top–right, bottom–left, bottom–right) on each trial. Because we were interested in location or object representation across condition (belief, desire, and think), we trained our pattern classifier to identify which of the four locations or objects was presented in each trial for two of the three conditions and then tested the classifier's performance in trials from the third condition, cross-validating across each possible permutation of training and test conditions. This allowed us to ensure that no single condition was driving the classifier's success. This strategy produced three separate object classification accuracy maps and three location classification accuracy maps for each participant. These classification maps were then corrected by subtracting chance accuracy (0.25) from each value so that resultant values reflected accuracy above chance. Because we were ultimately looking for regions of overlap between these separate classifications, the individual participant above-chance accuracy maps were smoothed by applying a 4.4 mm (2 voxel) Gaussian blur using AFNI's *3dMerge*. The smoothed accuracy maps were then warped to TLRC space using ANTs. Next, we ran one-sample *t* tests with AFNI's *3dttest++* to identify clusters of voxels where classification accuracy was significantly above chance across participants. Finally, we used the Monte Carlo simulation to perform cluster-wise corrections for multiple comparisons on the resulting *t* maps by running the *Clustsim* flag in *3dttest++*.

The second objective of our multivariate analyses was to identify distinct brain regions in which we could reliably decode the target object or target location within different attitude conditions. Although the brain must somehow distinguish believed propositions from non-believed propositions, the existence of attitude-specific object representations fits naturally with the Cartesian theory, according to which belief is a further process beyond comprehension. According to the Cartesian theory, a region might encode the identity of a hidden object as a dog, but only (or especially) when there is an active belief about the dog.

Likewise, regions might preferentially encode the identities of objects of desire or mere thought. We tested these hypotheses by searching for brain regions whose representational properties vary with propositional attitude.

To identify brain regions that preferentially encode object or location information only when participants are holding beliefs, desires, or merely thinking about the object or location, we used multivoxel pattern analysis to classify the target object or location on each trial within each condition. As with the previous analyses, the preprocessed BOLD signal for the 6–14 sec after the start of the critical 8-sec delay period was average to produce a single image for each trial. The data were then divided by condition, yielding 48 trials per condition, per participant. Using a whole-brain searchlight analysis and a leave-one-run-out cross-validation procedure, we trained our classifier to identify which of the four objects or which of the four locations was presented in each trial for 11 of 12 runs and then tested the classifier's performance in the 12th run, repeating this process for every permutation of training and test runs and averaging the classification accuracies across permutations. These analyses provided us with one set of whole-brain maps of the object classifier's accuracy for each condition for each participant, and a second similar set of maps for location classification accuracy. To assess differences in object or location representation between conditions, for each participant, we subtracted the voxel-wise accuracies for one condition from the voxel-wise accuracies for a second attitude condition (e.g., object classification accuracy in belief trials minus object classification accuracy in think trials). Both the within-condition accuracy maps and the maps of difference scores were warped to TLRC space using ANTs. Group-level significance testing was performed with one-sample *t* tests using AFNI's *3dttest++*, with whole-brain correction for multiple comparisons implemented using the *Clustsim* flag.

RESULTS

Behavioral Results

As expected, condition (belief vs. desire vs. think) had little effect on participants' responses to the attention check questions. To test the effect of question and condition on accuracy, for each participant, we computed an average accuracy for each question within each condition (e.g., "Is the [target object] in the [specific location]?" for belief trials, for desire trials, and for think trials separately). Across participants, the average accuracy for each question type within each condition was high, with a minimum average accuracy of 89.91% correct and a maximum average accuracy of 96.66% correct. We note that the correct response for some attention-check questions is "I don't know" (e.g., asking about the target object's location in "think" or "desire" trials). We then performed a mixed effect linear regression, implemented with the *lmer* function from the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015) to evaluate the degree to which question, condition, and their interaction predicted average response accuracy. This model also included a random intercept for each participant. We then compared this model to reduced models that excluded condition, question, or their interaction separately using the *ANOVA* function in R. We found that the full model performed significantly better than the model excluding the interaction term, $X^2(4, n = 70) = 16.85, p < .005$. This effect is likely driven by slightly reduced performance in response to the "Is the [target object] in the [location]?" question on belief trials (see Figure 2). It is worth noting, though, that accuracy in this case is still very high ($M = 89.92\%, SD = 0.13$). Critically, although there was some variation in accuracy across specific questions in specific conditions, accuracy was high for all questions in all conditions. Thus, our behavioral analyses confirmed that participants were consistently encoding the relevant information in all three conditions.

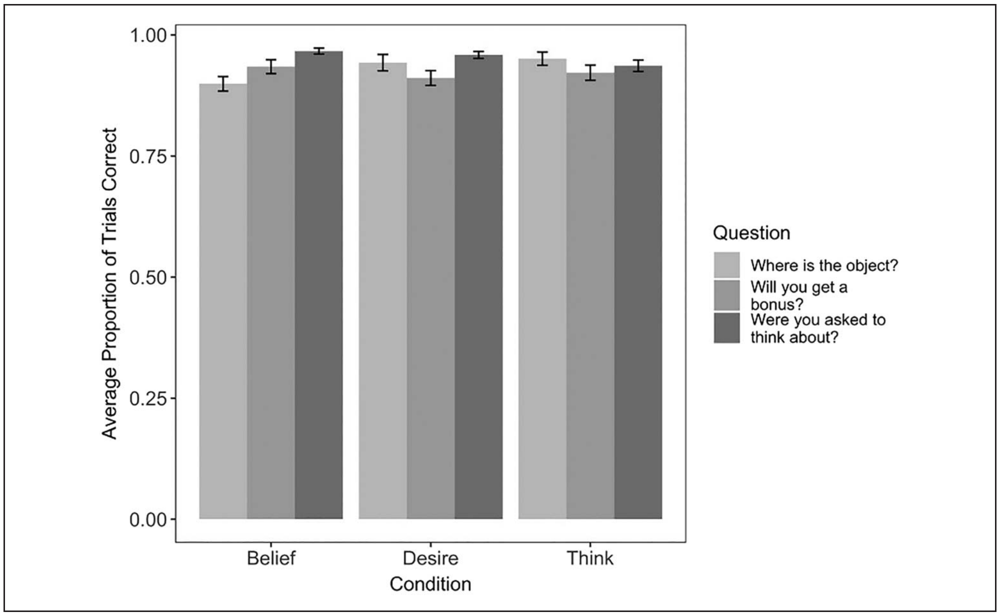


Figure 2. Behavioral results averaged across participants. After each trial, participants were asked one of three questions ("Is the [target object] in the [location]?", "Will you get a bonus if the [target object is in the [specific location]?", "Were you asked to think about the [target object] being in the [specific location]?"). One question was randomly selected for each trial. Participants were therefore asked each question in each of the three conditions.

Unbelieving

To directly test the Spinozan theory, we searched for regions exhibiting increased activity on trials when participants did not form a belief concerning the target object's location. That is, we focused on the contrasts *think* > *belief* and *desire* > *belief*. The *desire* > *belief* contrast yielded effects in 10 clusters exceeding the preregistered voxel-wise and cluster-wise corrected thresholds for significance. These include clusters in bilateral superior frontal gyrus, left supramarginal gyrus, middle temporal gyrus, and middle frontal gyrus (see Table 2). For the contrast *think* > *belief*, we observed effects in the right IFG and right insula at the prespecified voxel-wise threshold of $p < .001$. At this voxel-wise threshold, these clusters approach corrected significance at $p = .08$ and $p = .09$, respectively (two-tailed; see Table 1). We identified one 61 voxel cluster of overlap between these two contrasts (*desire* > *belief* and *think* > *belief*) in the right IFG (see Figure 3).

To further interrogate these overlapping results in the IFG, we reran these contrasts at a more stringent voxel-wise threshold of $p < .0005$. Using this higher voxel-wise threshold, both contrasts produced clusters in the right IFG that survive whole-brain correction for multiple comparisons at a cluster-wise corrected threshold of $p < .05$. We find a 35-voxel cluster overlap between the results of these two contrasts in the right IFG.

Believing

To directly test the Cartesian theory (which posits the engagement of a distinct belief process), we searched for regions exhibiting increased activity on trials when participants did form a belief concerning the target object's location. That is, we focused on the contrasts: *belief* > *desire* and *belief* > *think*. The *belief* > *desire* contrast yielded significant effects in the right precuneus,

Table 2. Univariate Activation Associated with *Non-belief*

		TLRC Coordinates[b]				
L/R	*Anatomical Region*[a]	*x*	*y*	*z*	*Peak z Score*	*Cluster Size (Voxels)*
Desire > *belief*						
L/R	Superior frontal gyrus	−11.5	42.5	32.5	6.83	10301***
L	Supramarginal gyrus	−47.5	−59.5	30.5	7.64	1776**
L	Middle temporal gyrus	−55.5	−31.5	−1.5	6.02	1565**
L	Middle frontal gyrus	−37.5	10.5	48.5	6.61	1020**
R	Middle temporal gyrus	56.5	−27.5	−9.5	5.41	735*
R	Supramarginal gyrus	48.5	−55.5	30.5	6.41	692*
L	IFG	−49.5	16.5	14.5	4.86	464*
L/R	Cingulate gyrus	−7.5	−19.5	30.5	5.45	396*
R	Pyramis	32.5	−73.5	33.5	5.14	377*
R	Middle frontal gyrus	40.5	−10.5	44.5	4.92	306*
Think > *belief*						
R	IFG	44.5	32.5	8.5	4.60	158[c]
R	Insula	40.5	4.5	14.5	4.25	369[c]
Overlap						
R	IFG	45.1	32.2	5.4	n/a	61

All clusters surpass a voxel-wise significance threshold of $p < .001$.

[a] Indicates anatomical region containing largest proportions of voxels, although in some cases cluster extends through additional regions.

[b] TLRC coordinates for voxels of peak activation for entire cluster. For overlap, it indicates coordinates in the approximate center of the cluster.

[c] Cluster-wise corrected $p < .10$.

* Cluster-wise corrected $p < .05$.

** Cluster-wise corrected $p < .01$.

*** Cluster-wise corrected $p < .005$.

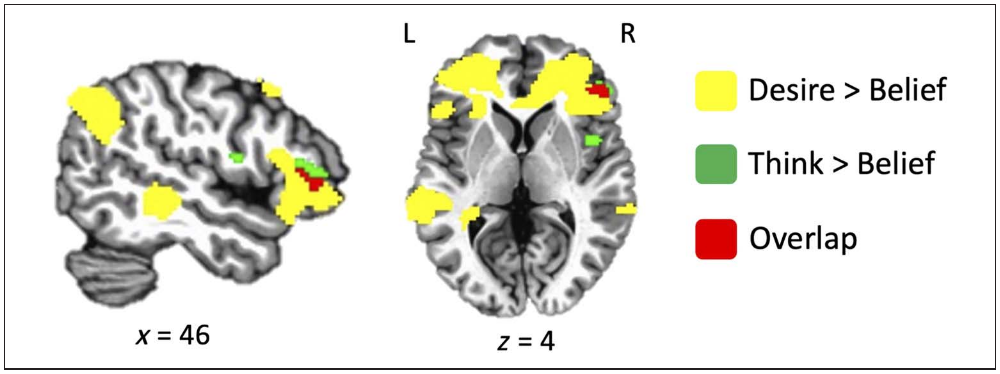


Figure 3. Greater neural activation associated with non-belief conditions relative to the belief condition in the right IFG. Desire > belief contrast: voxel-wise $p < .001$, cluster-wise corrected $p < .05$. Think > belief contrast: voxel-wise $p < .001$, cluster-wise corrected $p < .10$ (two-tailed). Coordinates are in TLRC space.

parahippocampal gyrus, middle occipital gyrus, and insula (voxel-wise $p < .001$, cluster-wise corrected $p < .05$). The *belief > think* contrast yielded significant effects in eight clusters including in the bilateral precuneus, bilateral middle frontal gyrus, and right thalamus (voxel-wise $p < .001$; cluster-wise corrected $p < .05$; see Table 3). We identified two clusters of overlap between these two contrasts (*belief > desire* and *belief > think*): a large cluster primarily in the right precuneus and a smaller cluster in the left cuneus (see Figure 4).

Table 3. Univariate Activation Associated with *Belief*

		TLRC Coordinates[b]				
L/R	*Anatomical Region*[a]	*x*	*y*	*z*	*Peak z Score*	*Cluster Size (Voxels)*
Belief > desire						
R	Precuneus	34.5	−81.5	26.5	7.09	11654***
R	Parahippocampal gyrus	40.5	−13.5	−23.5	5.68	1268**
L	Middle occipital gyrus	−33.5	−81.5	24.5	7.13	1237**
R	Insula	40.5	−9.5	−5.5	4.77	273*
Belief > think						
L/R	Precuneus	10.5	−73.5	52.5	6.74	6571***
R	Middle frontal gyrus	22.5	−5.5	48.5	5.58	1159**
L	Middle frontal gyrus	−19.5	−5.5	54.5	4.69	538*
R	Thalamus	16.5	−25.5	12.5	4.59	351*
L	Cerebellar tonsil	−15.5	−37.5	37.5	5.37	303*
R	Middle frontal gyrus	40.5	58.5	6.5	4.87	289*
L	Cerebellar tonsil	−41.5	−61.5	−37.5	4.63	288*
L	Cuneus	−27.5	−83.5	28.5	5.02	254*
Overlap						
R	Precuneus	19.0	−64.2	46.3	n/a	3542
L	Cuneus	−29.0	−82.0	27.9	n/a	188

All clusters surpass a voxel-wise significance threshold of $p < .001$.

[a] Indicates anatomical region containing largest proportions of voxels, although in some cases clusters extend through additional regions.

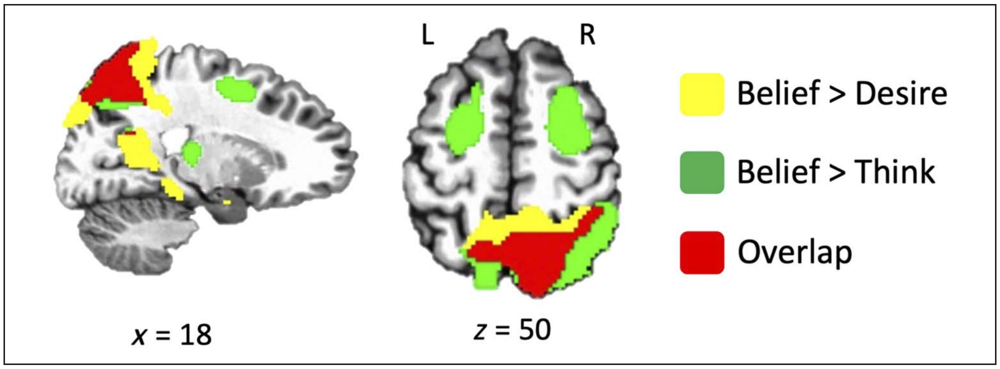
[b] TLRC coordinates for voxels of peak activation for entire cluster. For overlap, it indicates coordinates in the approximate center of the cluster.

* Cluster-wise corrected $p < .05$.

** Cluster-wise corrected $p < .01$.

*** Cluster-wise corrected $p < .005$.

Figure 4. Greater neural activation associated with the belief condition relative to the desire and think conditions in the right precuneus and left cuneus. Voxel-wise $p < .001$, cluster-wise corrected $p < .05$. Coordinates are in TLRC space.



Because *belief* trials also involved object tracking, whereas *desire* and *think* trials do not, we tested for a potential confound of object tracking by examining the neural activation within belief trials associated with each of our two shuffle types. Recall that the concealed objects were shuffled around the screen either by having the blue squares visibly rotate around the screen or by having arrows indicate the blue squares' final location. To test whether it is possible that the belief-specific activation identified in the previous analyses was related to object tracking, within each condition, we compared activation associated with each of the two shuffle types. Specifically, we used the same procedures for univariate analyses described above, but focused our analyses on the hemodynamic response corresponding to the 3-sec shuffle period. We then compared activation for trials that utilized the "rotation" shuffle to those that utilized the arrows shuffle. Across all three conditions, we found significantly increased activation for the rotation, when compared with the arrow shuffle, in the bilateral medial pFC (voxel-wise $p < .001$, cluster-wise $p < .05$; 2851 voxels; peak TLRC $x = -1.5$, $y = 26.5$, $z = -5.5$). Conversely, we found increased activation in the bilateral precuneus, inferior parietal lobule, and superior parietal lobe (voxel-wise $p < .001$, cluster-wise $p < .05$; 3790 voxels; peak TLRC $x = -23.5$, $y = -73.5$, $z = 44.5$) for the arrow shuffle when compared with the rotation. This cluster significantly overlaps with the cluster of voxels that were preferentially activated for *belief* rather than *desire* or *think* trials.

Desiring

To identify neural activity associated specifically with desiring, we performed the contrast *desire* > *think*, to complement the contrast *desire* > *belief* contrast performed above. The *desire* > *think* contrast yielded 12 clusters whose significance survived corrections for multiple comparisons (voxel-wise $p < .001$, cluster-wise corrected $p < .05$) including in the bilateral middle frontal gyrus, inferior parietal lobule, and middle temporal gyrus (see Table 4). Our overlap analysis revealed five distinct clusters of overlap, each greater than 100 voxels, in the posterior inferior parietal lobule, middle frontal gyrus, middle temporal gyrus, middle frontal gyrus, and dorsolateral pFC (see Figure 5).

Decoding Target Object/Location across Attitude Condition

As described above, to identify regions that reliably encode information about the target object or target location across conditions, we ran three separate whole-brain searchlight pattern-classification analyses—one with each condition held out as the test condition—seeking regions that support the decoding of the target object or target location regardless of which condition is withheld at test.

With the desire condition withheld at test, there was one cluster in which we were able to decode the identity of the target object above chance (voxel-wise $p < .005$, cluster-wise $p < .01$, 2686 voxels; centered at TLRC $x = 32.3$, $y = -1.7$, $z = 3.3$). This cluster was centered in the right caudate but included voxels in the right superior temporal gyrus, right lentiform nucleus, right insula, and right IFG. However, we were unable to decode the identity of the target object when either of the other two conditions were held out as the test condition.

Our three location classifications yielded a large region (13,102 voxels; centered at TLRC $x = 0.8$, $y = -76.4$, $z = 4.1$) encompassing most of the visual cortex in which we were able to decode the target location at significantly better than chance accuracy (voxel-wise $p < .005$, cluster-wise $p < .05$), regardless of which condition was held out as the test condition during classification.

Preferential Encoding of Object/Location within Attitude Condition

We were unable to decode the identity of the target object within condition at significantly better than chance accuracy. We were able to decode the identity of the target location above chance within each of the three conditions in the visual cortex (*belief*: 6965 voxels; peak TLRC $x = 2.5$, $y = -85.5$, $z = 4.5$; *desire*: 7496 voxels; peak TLRC $x = 8.5$, $y = -87.5$, $z = 6.5$; *think*: 6862 voxels; peak TLRC $x = 4.5$, $y = -75.5$, $z = 2.5$; for all three conditions voxel-wise $p < .005$, cluster-wise $p < .05$). However, we were unable to identify any clusters in which we were able to identify the target location significantly better in one condition when compared with either of the other two.

Table 4. Univariate Activation Associated with *Desire*

		TLRC Coordinates[b]				
L/R	*Anatomical Region*[a]	*x*	*y*	*z*	*Peak z Score*	*Cluster Size (Voxels)*
Desire > belief						
See Table 1						
Desire > think						
L	Middle frontal gyrus	−33.5	50.5	8.5	6.07	7221***
L	Inferior parietal lobule	−43.5	−55.5	30.5	7.50	2351***
R	Middle frontal gyrus	26.5	48.5	−2.5	6.22	1928***
R	Inferior parietal lobule	50.5	−57.5	44.5	6.35	1749**
R	Middle frontal gyrus	44.5	12.5	46.5	5.84	1632**
L	Middle temporal gyrus	−57.5	−29.5	−11.5	6.17	1191**
R	Pyramis	−7.5	−71.5	−25.5	5.39	1180**
R	Middle temporal gyrus	58.5	−29.5	−9.5	5.98	883*
L/R	Precuneus	−5.5	−69.5	34.5	5.40	759*
L	Cerebellar tonsil	−39.5	−57.5	−35.5	5.41	602*
L/R	Cingulate gyrus	−5.5	−23.5	30.5	4.59	372*
R	Caudate	8.5	8.5	8.5	5.78	350*
R	IFG	30.5	22.5	2.5	5.85	342*
Overlap						
L	Superior frontal gyrus	−7.7	36.5	32.1	n/a	3931
L	Inferior parietal lobule	−43.5	58	33.9	n/a	1526
R	Middle frontal gyrus	28.5	47.0	9.2		1219
L	Middle temporal gyrus	−55.6	−39.4	−5.0		1015
L	Middle frontal gyrus	−37.2	11.8	45.0		891
R	Supramarginal gyrus	47.3	−57.5	33.8		615
R	Middle temporal gyrus	58.1	−28.4	−7.4		584
R	Middle frontal gyrus	40.7	11.7	42.6		308
R	IFG	34.9	22.7	3.4		278
L/R	Cingulate	0.3	−23.8	30.3		241
L	IFG	−48.1	15.7	17.6		227
R	Pyramis	27.6	−68.3	−31.5		143

All clusters surpass a voxel-wise significance threshold of $p < .001$.

[a] Indicates anatomical region containing largest proportions of voxels, although in some cases cluster extends through additional regions.

[b] TLRC coordinates for voxels of peak activation for entire cluster. For overlap, it indicates coordinates in the approximate center of the cluster.

* Cluster-wise corrected $p < .05$.

** Cluster-wise corrected $p < .01$.

*** Cluster-wise corrected $p < .005$.

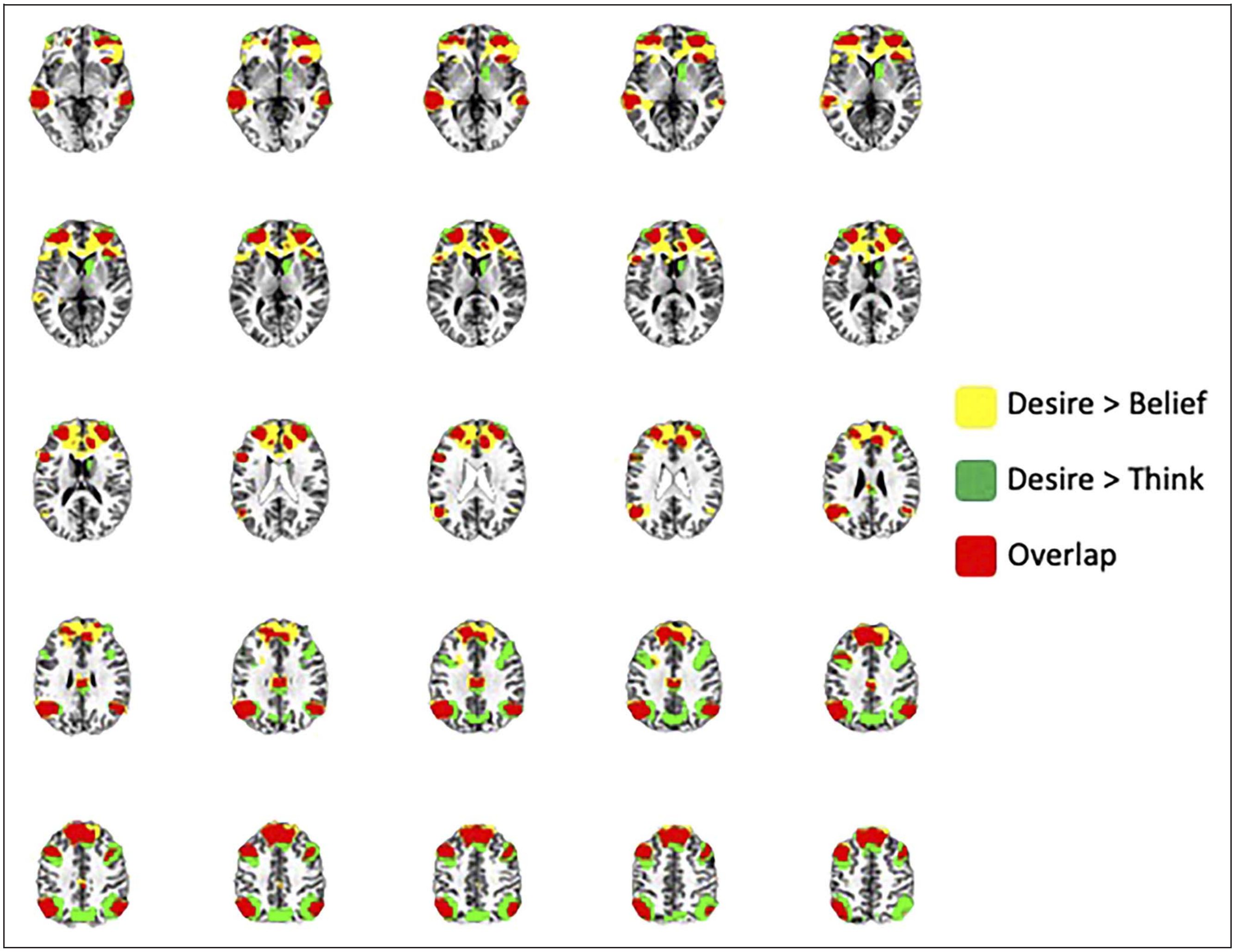


Figure 5. Greater neural activation associated with the desire condition relative to the belief and think conditions. Voxel-wise $p < .001$, cluster-wise corrected $p < .05$.

DISCUSSION

We provide evidence concerning a basic, unanswered question about the nature of belief: What distinguishes believing something from merely thinking about it? Prior behavioral research (Gilbert et al., 1990, 1993) has supported the Spinozan idea that thought without belief requires additional controlled processing. However, others have argued that these findings reflect uncertainty, rather than a process of unbelieving following a default tendency toward belief (Street & Kingstone, 2017; Street & Richardson, 2015). Neuroimaging studies have examined beliefs (Harris et al., 2008, 2009; Marques et al., 2009; Goel & Dolan, 2003), but they have not contrasted beliefs with other mental states with matched propositional/conceptual content. Nor have many of these studies distinguished metacognitive reporting on belief from belief itself. The large literature on confidence (degree of belief) in humans and animals is related to the present work (e.g., Platt & Huettel, 2008), but here, too, belief states are not contrasted with non-belief states matched for conceptual (or perceptual) content. Consequently, we have little empirical evidence concerning the cognitive and neural mechanisms that distinguish beliefs from other mental states with equivalent conceptual/propositional content. This is a significant gap in our understanding, as the capacity for entertaining complex ideas without believing in them appears to be an essential feature of high-level cognition, necessary for imagination, planning, and hypothesis testing in both science and everyday life.

Here, we deploy a new method, the shell game task, that enables us to experimentally induce states of belief, desire, and mere thought, while systematically controlling the propositional content of those mental states. Using this method, we have identified brain regions preferentially activated by believing, desiring, or merely thinking about ideas.

Most notably, we observed increased activity in the right IFG when participants represented an idea (or "proposition") without believing it. This was observed in both of the non-belief conditions (*think* > *belief* and *desire* >

belief). The increased activity in overlapping regions of the IFG for the two non-belief conditions provides evidence for the Spinozan theory of belief, which posits the existence of an additional process of "unbelieving" that is needed to entertain ideas without believing them. Moreover, these findings provide evidence for a specific version of the Spinozan theory whereby unbelieving is a process of inhibitory control supported by the IFG. This interpretation depends on a "reverse inference" (Poldrack, 2006), but one that is well supported by the literature.

Right IFG functioning is relatively heterogeneous. The region plays an important role in phonological and prosodic processing (Hartwigsen et al., 2010; Rota et al., 2009) and for the recovery of language function from stroke-induced aphasia (van Oers et al., 2010; Winhuisen et al., 2005). The right IFG has also been implicated in social cognition (Kilner, Neal, Weiskopf, Friston, & Frith, 2009), playing an important role in the experience of empathy (Shamay-Tsoory, Aharon-Peretz, & Perry, 2009), inferring unseen actions (Umilta et al., 2001), and theory of mind (Samson, Houthuys, & Humphreys, 2015; Samson, Apperly, & Humphreys, 2007).

However, this region has been most consistently implicated in response inhibition (see Aron, Robbins, & Poldrack, 2004, 2014, for reviews). The majority of the research on the right IFG and response inhibition has used the go/no-go and stop-signal tasks (see Swick, Ashley, & Turken, 2011, for a meta-analysis). In one such relevant study, the authors used a modified go/no-go task to differentiate between the cognitive inhibition, motor inhibition, and action cancellation components of motor inhibition tasks (Sebastian et al., 2013). Although engaging in inhibition in general increased activation in large portions of the right IFG, only the cognitive inhibition components of the task increased activation in the portion of the right IFG preferentially activated by non-belief trials in our study.

The right IFG has also been found to play a significant role in other types of psychological inhibition. It is preferentially engaged during some types of emotional inhibition including voluntarily suppressing negative affect (Phan et al., 2005), engaging in reappraisal to reduce emotional distress (Kim & Hamann, 2007; Ochsner et al., 2004), and mitigating emotional distractions (Dolcos & McCarthy, 2006). It has also been implicated in tasks requiring cognitive control, such as intentionally suppressing thoughts or memories as well as resolving stimulus conflict (Mitchell et al., 2007; Anderson & Green, 2001; Egner, 2011).

Finally, the IFG has been implicated specifically in withholding belief. Activation in the right IFG is increased when participants are required to withhold belief to evaluate the validity of logical syllogisms (Goel & Dolan, 2003), and disruption of the right IFG with TMS interferes with their ability to do so (Tsuji et al., 2010, 2011). Damage to the vmPFC also makes individuals increasingly susceptible to believing misleading advertising, even when they are told that the content of the advertisements is false (Asp et al., 2012). In light of such findings, Asp and Tranel (2013) have argued that the pFC is critical to Spinozan unbelieving, although they suggest that PFC serves to tag false beliefs rather than as engaging the inhibitory process we posit here.

The well-established importance of the right IFG in response inhibition, and the fact that our specific ROI has been implicated in several studies involving inhibition and cognitive control, suggests that it may also play an inhibitory role in our task. Importantly, the effects observed in our study create a double dissociation with a distinct set of brain regions that are more active for belief. Consequently, the effects associated with non-belief cannot be accounted for as products of increased overall attention or engagement. The desire and think conditions both involve the presentation of an additional verbal cue related to the target location, raising the possibility that the observed effects in the IFG reflect linguistic processing spilling over into the critical delay period. However, these effects are right-lateralized, the opposite of what one would predict on this alternative explanation. We also note that we observed no effects classically associated with reading (e.g., in the visual word form area) for the desire and think conditions during the critical delay period.

Other studies implicate the IFG in thought without belief. TMS to the left IFG has been shown to decrease susceptibility to the "good news/bad news" effect (Sharot et al., 2012), in which individuals more readily incorporate positive news into their beliefs but show an aversion to incorporating negative news. In this work, participants who received real TMS were more likely to believe bad news than those who received sham TMS. Finally, very recent research shows that the right IFG is preferentially engaged during the consideration or evaluation of counterfactual events (those that could have happened but did not), but not of events that actually occurred (Bernhard, Cushman, & Phillips, in prep; Nieuwland, 2012). In conjunction, this body of evidence, along with the current results, suggests a role for the IFG in inhibiting belief.

The increased activation in the IFG for desiring over believing is also consistent with an account of "wishful thinking" or "desirability bias" (Windschitl, Smith, Rose, & Krizan, 2010; Krizan & Windschitl, 2009) as a special case of Spinozan belief bias: Desiring a state of the world inherently involves thinking about that state of the world, which produces a belief in that state of the world unless it is inhibited by a process of unbelieving (Mandelbaum, 2014). This is consistent with decreased wishful thinking (Bamford & Lagattuta, 2020) and increased inhibitory control (Christ, White, Mandernach, & Keys, 2001) in children as they mature. We note, however, that desiring, relative to thinking, was associated with increased activity in a broad set of pFC regions, many of which are not specifically associated with inhibitory control. This suggests the need for further investigation specifically aimed at testing the hypothesis that wishful thinking results from a failure of unbelieving, understood as a failure of inhibitory control.

Although we interpret our findings in the right IFG as evidence of Spinozan inhibition, one might wonder whether our findings of increased activity for belief (relative to both desire and mere thought) in the cuneus and precuneus provide at least some support for the Cartesian theory, which posits that belief as a further process beyond comprehension. To the extent that this finding supports a Cartesian interpretation, it is a nonstandard Cartesian theory. The standard Cartesian theory posits a distinct process of truth-evaluation following comprehension, but it does not posit a distinct process that is specific to belief. Accepting our interpretation of our primary findings for the IFG, this would need to be a hybrid Spinozan–Cartesian theory whereby belief is a distinct process (Cartesian) that coexists with, and may be inhibited by, a Spinozan inhibitory process. Although this is an intriguing possibility, it depends on the assumption that the observed activity in the cuneus and precuneus do in fact reflect belief-specific processes. Further analysis casts doubt on this, as these effects may be because of object-tracking processes spilling over into the delay period, despite our efforts to equate the perceptual properties of each condition during the delay period. Anticipating the possibility of unwanted perceptual spillover from the tracking phase of each trial, we used two different cues for location change during the shuffling period (Figure 1), arrows and actual motion. We found stronger effects in these regions for trials involving arrows, rather than actual motion. The cuneus is a classic visual region in the occipital lobe (e.g., Vanni, Tanskanen, Seppä, Uutela, & Hari, 2001), and the subregion of the precuneus identified here appears to play an important role in object tracking (Kimmig et al., 2008; Shulman et al., 1999). In light of the above, we suggest that these belief-related results may be artifactual and require further investigation. We note, however, that we found no effect of cue type in the IFG for any analysis.

We also observed many regions that were preferentially activated for desire trials (*desire* > *belief* and *desire* > *think*). Given that desire in our task is elicited by the anticipation of a rewarding outcome, one might expect desire trials to elicit increased activity in regions associated with the anticipation of reward. However, we found no evidence of increased activation in classic reward-related regions such as the ventral striatum and orbital frontal cortex. This may be because we conducted our analyses during an extended delay period after the presentation of reward-relevant information. Instead, many of the regions that exhibited increased activity for desire trials lie within the default mode network (Raichle, 2015; Buckner, Andrews-Hanna, & Schacter, 2008). These include regions in the medial pFC, the inferior parietal lobule, and the middle temporal gyrus. We likewise observed desire-related activity in the dorsal attention network (Szczepanski, Pinsk, Douglas, Kastner, & Saalmann, 2013; Fox, Corbetta, Snyder, Vincent, & Raichle, 2006; Corbetta & Shulman, 2002), including subregions of the dorsolateral pFC. Although activity within these two networks tends to be anticorrelated (Fox, Zhang, Snyder, & Raichle, 2009), they can simultaneously exhibit increased activation during mental simulations of future goal-directed action (Stawarczyk & D'Argembeau, 2015; Gerlach, Spreng, Madore, & Schacter, 2014; Gerlach, Spreng, Gilmore, & Schacter, 2011). Although participants could not take action to increase their probability of reward in this task, desire may naturally elicit thoughts about future behavior. Consistent with this idea, many of these regions have also been implicated in the valuation of imagined possibilities (Bulley, Henry, & Suddendorf, 2016; Benoit, Szpunar, & Schacter, 2014), and episodic simulation more generally (De Brigard, Addis, Ford, Schacter, & Giovanello, 2013; Schacter et al., 2012; Addis, Pan, Vu, Laiser, & Schacter, 2009).

Finally, our multivariate analyses yielded only one significant finding: target location encoding across attitude conditions in the visual cortex. This most likely reflects a shift of attention to the target location. We were unable to reliably decode the identity of the target objects across conditions, or within condition. The target object is not visible to participants during the critical delay period in our task, making it unlikely that object identity would be decodable in visual regions. Some studies report decoding object identity from visual working memory (e.g., Emrich, Riggall, LaRocque, & Postle, 2013; Riggall & Postle, 2012; Christophel, Hebart, & Haynes, 2012; Harrison & Tong, 2009), but others failed to do so (Olmos-Solis, van Loon, & Olivers, 2021; Linden, Oosterhof, Klein, & Downing, 2012).

Here, we introduced a new task, the shell game, to examine the neural mechanisms of belief, desire, and mere thought in a controlled fashion. This task, however, is somewhat artificial and focuses on propositional content related to concrete states of affairs (objects and their locations) with information presented visually and (primarily) nonverbally. It is unknown whether the present results will generalize to verbal presentation of information and/or to information presented in other sensory modalities (e.g., auditory). Likewise, it is unknown whether these results will generalize to propositional attitudes related to the social domain (e.g., "Caitlin is married to Elizabeth," "James has three children") and to abstract propositions more generally (e.g., "Inflation is rising in the United States," "God exists"). Although propositional attitudes toward many such propositions would be difficult to control experimentally, our experimental paradigm could be adapted to examine beliefs, desire, and mere thoughts about more abstract states of affairs and to propositional attitudes generated by verbal stimuli. We view these as important directions for future research.

Here, we have presented evidence for the Spinozan account of belief (Gilbert et al., 1993; Gilbert, 1991; Spinoza, 1677/1982). Although the Spinozan view predates Darwin, it has a natural evolutionary explanation. Automatically believing information presented by the environment will be adaptive in an environment in which most of the available information is accurate (Mandelbaum & Quilty-Dunn, 2015; Levine, 2014; Kissine & Klein, 2013; Reber & Unkelbach, 2010; Levine et al., 1999). Before the advent of language, when beliefs were based entirely on typically veridical perceptions, this assumption may have held (Mandelbaum & Quilty-Dunn, 2015; Mandelbaum, 2014). However, the capacity for language opens human minds to a vast sea of ideas—some true, some false. Humans have the useful ability to form accurate beliefs about things far beyond their personal experiences, but humans are also highly susceptible to misinformation, whether intentional or unintentional (Van Der Linden, 2022). This creates a need for a cognitive "spam filter," a mechanism for entertaining ideas without believing them.

If humans have a unique (or uniquely prominent) need for "unbelieving" thanks to language, this naturally raises questions about what happens when the unbelieving process is compromised. Schizophrenia and related mental disorders involve delusions, beliefs that are firmly held despite disconfirmatory evidence (American Psychiatric Association, 2013). We suggest that delusion may be understood as a failure of Spinozan unbelieving, which we propose may involve a failure of inhibitory control. It is likely that delusions are not merely the result of failed inhibitory control, as many disorders involving failures of inhibitory control do not involve delusions (e.g., attention deficit hyperactivity disorder; Quay, 1997). However other evidence suggests that delusions have an inhibitory component, as patients with schizophrenia exhibit accelerated semantic spreading (Kreher, Goff, & Kuperberg, 2009; Moritz, Woodward, Küppers, Lausen, & Schickel, 2003; Spitzer, Braun, Hermle, & Maier, 1993; Maher, Manschreck, Hoover, & Weisstein, 1987) and a tendency to jump to unwarranted conclusions (Dudley et al., 2016; Evans et al., 2015; Moritz & Woodward, 2005). Delusions are difficult to study because the formation of delusional belief typically occurs outside the laboratory and without control of the available information. We suggest that the shell game paradigm may be used to study the process of belief formation in clinical populations and to test the hypothesis that delusions result from a failure of Spinozan unbelieving. More generally, however, the shell game paradigm provides a method for studying the mechanisms of propositional attitudes in controlled fashion, with the ability to dissociate propositional attitudes from propositional content.

Reprint requests should be sent to Regan Bernhard, Boston College, Department of Psychology and Neuroscience, McGuinn 430A, 275 Beacon St., Chestnut Hill, MA 02467, or e-mail: regan.bernhard@gmail.com.

Data Availability Statement

Imaging data are available upon request from the corresponding author.

Author Contributions

Regan Bernhard: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Visualization; Writing—Original draft; Writing—Review & editing. Steven Frankland: Conceptualization; Formal analysis; Investigation; Methodology; Software; Writing—Review & editing. Dillon Plunkett: Data collection and analyses; Revisions; Writing. Beau Sievers: Formal analysis; Writing—Review & editing. Joshua Greene: Conceptualization; Investigation; Methodology; Supervision; Visualization; Writing—Original draft; Writing—Review & editing.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .62; W/M = .26; M/W = .075; W/W = .075.

REFERENCES

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage*, *31*, 790–795. https://doi.org/10.1016/j.neuroimage.2006.01.001, PubMed: 16487726

Addis, D. R., Pan, L., Vu, M. A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, *47*, 2222–2238. https://doi.org/10.1016/j.neuropsychologia.2008.10.026, PubMed: 19041331

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Anderson, C. A. (1983). Abstract and concrete data in the perseverance of social theories: When weak data lead to unshakeable beliefs. *Journal of Experimental Social Psychology*, *19*, 93–108. https://doi.org/10.1016/0022-1031(83)90031-8

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, *410*, 366–369. https://doi.org/10.1038/35066572, PubMed: 11268212

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*, 1037–1049. https://doi.org/10.1037/h0077720

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J., & Poldrack, R. A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience*, *27*, 3743–3752. https://doi.org/10.1523/JNEUROSCI.0519-07.2007, PubMed: 17409238

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*, 170–177. https://doi.org/10.1016/j.tics.2004.02.010, PubMed: 15050513

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*, 177–185. https://doi.org/10.1016/j.tics.2013.12.003, PubMed: 24440116

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*, 100. https://doi.org/10.3389/fnins.2012.00100, PubMed: 22787439

Asp, E., & Tranel, D. (2013). False tagging theory: Toward a unitary account of prefrontal cortex function. In D. T. Stuss, & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 383–416). Oxford University Press. https://doi.org/10.1093/med/9780199837755.003.0029

Aue, T., Nusbaum, H. C., & Cacioppo, J. T. (2012). Neural correlates of wishful thinking. *Social Cognitive and Affective Neuroscience*, *7*, 991–1000. https://doi.org/10.1093/scan/nsr081, PubMed: 22198967

Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). *Insight Journal*, *2*, 1–35. https://doi.org/10.54294/uvnhin

Babad, E. (1997). Wishful thinking among voters: Motivational and cognitive influences. *International Journal of Public Opinion Research*, *9*, 105–125. https://doi.org/10.1093/ijpor/9.2.105

Bamford, C., & Lagattuta, K. H. (2020). Optimism and wishful thinking: Consistency across populations in children's expectations for the future. *Child Development*, *91*, 1116–1134. https://doi.org/10.1111/cdev.13293, PubMed: 31418461

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bear, A., Fortgang, R. G., Bronstein, M. V., & Cannon, T. D. (2017). Mistiming of thought and perception predicts delusionality. *Proceedings of the National Academy of Sciences, U.S.A.*, *114*, 10791–10796. https://doi.org/10.1073/pnas.1711383114, PubMed: 28923963

Begg, I., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*, 446–458. https://doi.org/10.1037/0096-3445.121.4.446

Benoit, R. G., Szpunar, K. K., & Schacter, D. L. (2014). Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, 16550–16555. https://doi.org/10.1073/pnas.1419274111, PubMed: 25368170

Bernhard, R. M., Cushman, F., & Phillips, J. (in prep). The neural instantiation of counterfactual thought.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234. https://doi.org/10.1207/s15327957pspr1003_2, PubMed: 16859438

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*, 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences*, *1124*, 1–38. https://doi.org/10.1196/annals.1440.011, PubMed: 18400922

Bulley, A., Henry, J., & Suddendorf, T. (2016). Prospection and the present moment: The role of episodic foresight in intertemporal choices between immediate and delayed rewards. *Review of General Psychology*, *20*, 29–47. https://doi.org/10.1037/gpr0000061

Cahill, D. P. (2015). *Wishful thinking, fast and slow* (Doctoral dissertation).

Chein, J. M., & Schneider, W. (2005). Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. *Cognitive Brain Research*, *25*, 607–623. https://doi.org/10.1016/j.cogbrainres.2005.08.013, PubMed: 16242923

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, *29*, 12315–12320. https://doi.org/10.1523/JNEUROSCI.2575-09.2009, PubMed: 19793990

Christ, S. E., White, D. A., Mandernach, T., & Keys, B. A. (2001). Inhibitory control across the life span. *Developmental Neuropsychology*, *20*, 653–669. https://doi.org/10.1207/S15326942DN2003_7, PubMed: 12002099

Christophel, T. B., Hebart, M. N., & Haynes, J. D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, *32*, 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012, PubMed: 22993415

Cole, M. W., & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *Neuroimage*, *37*, 343–360. https://doi.org/10.1016/j.neuroimage.2007.03.071, PubMed: 17553704

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*, 201–215. https://doi.org/10.1038/nrn755, PubMed: 11994752

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron*, *29*, 537–545. https://doi.org/10.1016/S0896-6273(01)00225-2, PubMed: 11239442

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, *51*, 2401–2414. https://doi.org/10.1016/j.neuropsychologia.2013.01.015, PubMed: 23376052

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *Neuroscientist*, *13*, 580–593. https://doi.org/10.1177/1073858407304654, PubMed: 17911216

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*, 238–257. https://doi.org/10.1177/1088868309352251, PubMed: 20023210

Delgado, M. R., Miller, M. M., Inati, S., & Phelps, E. A. (2005). An fMRI study of reward-related probability learning. *Neuroimage*, *24*, 862–873. https://doi.org/10.1016/j.neuroimage.2004.10.002, PubMed: 15652321

Descartes, R. (1984). Principles of philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Eds. and Trans.), *The philosophical writings of Descartes* (Vol. 1, pp. 193–291). Cambridge, UK: Cambridge University Press. (Original work published 1644). https://doi.org/10.1017/CBO9780511805042.007

Dias, R., Robbins, T. W., & Roberts, A. C. (1997). Dissociable forms of inhibitory control within prefrontal cortex with an analog of the Wisconsin Card Sort Test: Restriction to novel situations and independence from "on-line" processing. *Journal of Neuroscience*, *17*, 9285–9297. https://doi.org/10.1523/JNEUROSCI.17-23-09285.1997, PubMed: 9364074

Dolcos, F., & McCarthy, G. (2006). Brain systems mediating cognitive interference by emotional distraction. *Journal of Neuroscience*, *26*, 2072–2079. https://doi.org/10.1523/JNEUROSCI.5042-05.2006, PubMed: 16481440

Dreher, J. C., Kohn, P., & Berman, K. F. (2006). Neural coding of distinct statistical properties of reward information in humans. *Cerebral Cortex*, *16*, 561–573. https://doi.org/10.1093/cercor/bhj004, PubMed: 16033924

Dudley, R., Taylor, P., Wickham, S., & Hutton, P. (2016). Psychosis, delusions and the "jumping to conclusions" reasoning bias: A systematic review and meta-analysis. *Schizophrenia Bulletin*, *42*, 652–665. https://doi.org/10.1093/schbul/sbv150, PubMed: 26519952

Egner, T. (2011). Right ventrolateral prefrontal cortex mediates individual differences in conflict-driven cognitive control. *Journal of Cognitive Neuroscience*, *23*, 3903–3913. https://doi.org/10.1162/jocn_a_00064, PubMed: 21568631

Emrich, S. M., Riggall, A. C., LaRocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual shortterm memory. *Journal of Neuroscience*, *33*, 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013, PubMed: 23575849

Evans, S. L., Averbeck, B. B., & Furl, N. (2015). Jumping to conclusions in schizophrenia. *Neuropsychiatric Disease and Treatment*, *11*, 1615–1624. https://doi.org/10.2147/NDT.S56870, PubMed: 26170674

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002. https://doi.org/10.1037/xge0000098, PubMed: 26301795

Fiedler, K., Armbruster, T., Nickel, S., Walther, E., & Asbeck, J. (1996). Constructive biases in social judgment: Experiments on the self-verification of question contents. *Journal of Personality and Social Psychology*, *71*, 861–873. https://doi.org/10.1037/0022-3514.71.5.861, PubMed: 8939037

Fiedler, K., Walther, E., Armbruster, T., Fay, D., & Naumann, U. (1996). Do you really know what you have seen? Intrusion errors and presuppositions effects on constructive memory. *Journal of Experimental Social Psychology*, *32*, 484–511. https://doi.org/10.1006/jesp.1996.0022

Fillenbaum, S. (1966). Memory for gist: Some relevant variables. *Language and Speech*, *9*, 217–227. https://doi.org/10.1177/002383096600900403, PubMed: 5975860

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902. https://doi.org/10.1126/science.1077349, PubMed: 12649484

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences, U.S.A.*, *103*, 10046–10051. https://doi.org/10.1073/pnas.0604187103, PubMed: 16788060

Fox, M. D., Zhang, D., Snyder, A. Z., & Raichle, M. E. (2009). The global signal and observed anticorrelated resting state brain networks. *Journal of Neurophysiology*, *101*, 3270–3283. https://doi.org/10.1152/jn.90777.2008, PubMed: 19339462

Gerlach, K. D., Spreng, R. N., Gilmore, A. W., & Schacter, D. L. (2011). Solving future problems: Default network and executive activity associated with goal-directed mental simulations. *Neuroimage*, *55*, 1816–1824. https://doi.org/10.1016/j.neuroimage.2011.01.030, PubMed: 21256228

Gerlach, K. D., Spreng, R. N., Madore, K. P., & Schacter, D. L. (2014). Future planning: Default network activity couples with frontoparietal control network and reward-processing regions during process and outcome simulations. *Social Cognitive and Affective Neuroscience*, *9*, 1942–1951. https://doi.org/10.1093/scan/nsu001, PubMed: 24493844

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

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*, 601–613. https://doi.org/10.1037/0022-3514.59.4.601

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*, 221–233. https://doi.org/10.1037/0022-3514.65.2.221, PubMed: 8366418

Glimcher, P. (2003). *Decisions, uncertainty and the brain: The science of neuroecomnomics*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/2302.001.0001

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, *87*, B11–B22. https://doi.org/10.1016/S0010-0277(02)00185-3, PubMed: 12499108

Guenther, C. L., & Alicke, M. D. (2008). Self-enhancement and belief perseverance. *Journal of Experimental Social Psychology*, *44*, 706–712. https://doi.org/10.1016/j.jesp.2007.04.010

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences, U.S.A.*, *98*, 4259–4264. https://doi.org/10.1073/pnas.071043098, PubMed: 11259662

Hampton, A. N., & O'doherty, J. P. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. *Proceedings of the National Academy of sciences, U.S.A.*, *104*, 1377–1382. https://doi.org/10.1073/pnas.0606297104, PubMed: 17227855

Hare, T. A., O'doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, *28*, 5623–5630. https://doi.org/10.1523/JNEUROSCI.1309-08.2008, PubMed: 18509023

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*, e7272. https://doi.org/10.1371/journal.pone.0007272, PubMed: 19794914

Harris, S., Sheth, S. A., & Cohen, M. S. (2008). Functional neuroimaging of belief, disbelief, and uncertainty. *Annals of Neurology*, *63*, 141–147. https://doi.org/10.1002/ana.21301, PubMed: 18072236

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*, 632–635. https://doi.org/10.1038/nature07832, PubMed: 19225460

Hartwigsen, G., Price, C. J., Baumgaertner, A., Geiss, G., Koehnke, M., Ulmer, S., et al. (2010). The right posterior inferior frontal gyrus contributes to phonological word decisions in the healthy brain: Evidence from dual-site TMS. *Neuropsychologia*, *48*, 3155–3163. https://doi.org/10.1016/j.neuropsychologia.2010.06.032, PubMed: 20600177

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe it or not: On the possibility of suspending belief. *Psychological Science*, *16*, 566–571. https://doi.org/10.1111/j.0956-7976.2005.01576.x, PubMed: 16008791

Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, *19*, 212–225. https://doi.org/10.1086/209297

Heekeren, H. R., Marrett, S., Bandettini, P. A., & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, *431*, 859–862. https://doi.org/10.1038/nature02966, PubMed: 15483614

Henkel, L. A., & Mattson, M. E. (2011). Reading is believing: The truth effect and source credibility. *Consciousness and Cognition*, *20*, 1705–1721. https://doi.org/10.1016/j.concog.2011.08.018, PubMed: 21978908

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*, 1625–1633. https://doi.org/10.1038/nn2007, PubMed: 17982449

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*, 227–231. https://doi.org/10.1038/nature07200, PubMed: 18690210

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*, 759–764. https://doi.org/10.1126/science.1169405, PubMed: 19423820

Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *Journal of Neuroscience*, *29*, 10153–10159. https://doi.org/10.1523/JNEUROSCI.2668-09.2009, PubMed: 19675249

Kim, S. H., & Hamann, S. (2007). Neural correlates of positive and negative emotion regulation. *Journal of Cognitive Neuroscience*, *19*, 776–798. https://doi.org/10.1162/jocn.2007.19.5.776, PubMed: 17488204

Kimmig, H., Ohlendorf, S., Speck, O., Sprenger, A., Rutschmann, R., Haller, S., et al. (2008). fMRI evidence for sensorimotor transformations in human cortex during smooth pursuit eye movements. *Neuropsychologia*, *46*, 2203–2213. https://doi.org/10.1016/j.neuropsychologia.2008.02.021, PubMed: 18394660

Kissine, M., & Klein, O. (2013). Models of communication, epistemic trust and epistemic vigilance. In J. Laszlo, J. Forgas, & O. Vincze (Eds.), *Social cognition and communication*. New York: Psychology Press.

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: Neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *363*, 3771–3786. https://doi.org/10.1098/rstb.2008.0155, PubMed: 18829428

Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007). Neural predictors of purchases. *Neuron*, *53*, 147–156. https://doi.org/10.1016/j.neuron.2006.11.010, PubMed: 17196537

Kreher, D. A., Goff, D., & Kuperberg, G. R. (2009). Why all the confusion? Experimental task explains discrepant semantic priming effects in schizophrenia under “automatic” conditions: Evidence from event-related potentials. *Schizophrenia Research*, *111*, 174–181. https://doi.org/10.1016/j.schres.2009.03.013, PubMed: 19386472

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage*, *38*, 649–662. https://doi.org/10.1016/j.neuroimage.2007.02.022, PubMed: 17804260

Krizan, Z., & Windschitl, P. D. (2009). Wishful thinking about the future: Does desire impact optimism? *Social and Personality Psychology Compass*, *3*, 227–243. https://doi.org/10.1111/j.1751-9004.2009.00169.x

Levine, T. R. (2014). Truth-default theory (TDT). *Journal of Language and Social Psychology*, *33*, 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the “veracity effect”. *Communications Monographs*, *66*, 125–144. https://doi.org/10.1080/03637759909376468

Linden, D. E., Oosterhof, N. N., Klein, C., & Downing, P. E. (2012). Mapping brain activation and information during category-specific visual working memory. *Journal of Neurophysiology*, *107*, 628–639. https://doi.org/10.1152/jn.00105.2011, PubMed: 22013235

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220. https://doi.org/10.1146/annurev-neuro-071013-014017, PubMed: 25032495

Maher, B., Manschreck, T. C., Hoover, T. M., & Weisstein, C. C. (1987). Thought disorder and measured features of language production in schizophrenia. In P. D. Harvey & E. E. Walker (Eds.), *Positive and negative symptoms of psychosis: Description, research, and future directions* (pp. 195–215). Lawrence Erlbaum Associates.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry*, *57*, 55–96. https://doi.org/10.1080/0020174X.2014.858417

Mandelbaum, E., & Quilty-Dunn, J. (2015). Believing without reason, or: Why liberals shouldn’t watch Fox News. *The Harvard Review of Philosophy*, *22*, 42–52. https://doi.org/10.5840/harvardreview2015226

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*, 759–768. https://doi.org/10.1016/j.cortex.2008.07.004, PubMed: 19059586

Mazoyer, B., Zago, L., Mellet, E., Bricogne, S., Etard, O., Houdé, O., et al. (2001). Cortical networks for working memory and executive functions sustain the conscious resting state in man. *Brain Research Bulletin*, *54*, 287–298. https://doi.org/10.1016/S0361-9230(00)00437-8, PubMed: 11287133

Mazzoni, G., & Nelson, T. O. (Eds.). (1998). *Metacognition and cognitive neuropsychology*. Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781315805733

McKay, T., & Nelson, M. (2014). Propositional attitude reports. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2014 ed.). Metaphysics Research Lab, Stanford University.

Metcalfe, J., & Shimamura, A. P. (Eds.). (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4561.001.0001

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*, 78–92. https://doi.org/10.1016/j.neuron.2015.09.039, PubMed: 26447574

Mitchell, J. P., Heatherton, T. F., Kelley, W. M., Wyland, C. L., Wegner, D. M., & Macrae, C. N. (2007). Separating sustained from transient aspects of cognitive control during thought suppression. *Psychological Science*, *18*, 292–297. https://doi.org/10.1111/j.1467-9280.2007.01891.x, PubMed: 17470250

Moritz, S., & Woodward, T. S. (2005). Jumping to conclusions in delusional and non-delusional schizophrenic patients. *British Journal of Clinical Psychology*, *44*, 193–207. https://doi.org/10.1348/014466505X35678, PubMed: 16004654

Moritz, S., Woodward, T. S., Küppers, D., Lausen, A., & Schickel, M. (2003). Increased automatic spreading of activation in thought-disordered schizophrenic patients. *Schizophrenia Research*, *59*, 181–186. https://doi.org/10.1016/S0920-9964(01)00337-1, PubMed: 12414074

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O’Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences*, *15*, 453–459. https://doi.org/10.1016/j.tics.2011.07.011, PubMed: 21889391

Nadarevic, L., & Erdfelder, E. (2013). Spinoza’s error: Memory for truth and falsity. *Memory & Cognition*, *41*, 176–186. https://doi.org/10.3758/s13421-012-0251-z, PubMed: 22972664

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience*, *12*, 241–268. https://doi.org/10.3758/s13415-011-0083-5, PubMed: 22282036

Nieuwland, M. S. (2012). Establishing propositional truth-value in counterfactual and real-world contexts during sentence comprehension: Differential sensitivity of the left and right inferior frontal gyri. *Neuroimage*, *59*, 3433–3440. https://doi.org/10.1016/j.neuroimage.2011.11.018, PubMed: 22116039

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., et al. (2004). For better or for worse: Neural systems supporting the cognitive down-and up-regulation of negative emotion. *Neuroimage*, *23*, 483–499. https://doi.org/10.1016/j.neuroimage.2004.06.030, PubMed: 15488398

Olmos-Solis, K., van Loon, A. M., & Olivers, C. N. (2021). Content or status: Frontal and posterior cortical representations of object category and upcoming task goals in working memory. *Cortex*, *135*, 61–77. https://doi.org/10.1016/j.cortex.2020.11.011, PubMed: 33360761

O'Neill, M., & Schultz, W. (2010). Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron*, *68*, 789–800. https://doi.org/10.1016/j.neuron.2010.09.031, PubMed: 21092866

Pantazi, M., Kissine, M., & Klein, O. (2018). The power of the truth bias: False information affects memory and judgment even in the absence of distraction. *Social Cognition*, *36*, 167–198. https://doi.org/10.1521/soco.2018.36.2.167

Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: A comparative study. *Neuroimage*, *56*, 476–496. https://doi.org/10.1016/j.neuroimage.2010.05.026, PubMed: 20488249

Peter, C., & Koch, T. (2016). When debunking scientific myths fails (and when it does not): The backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication*, *38*, 3–25. https://doi.org/10.1177/1075547015613523

Phan, K. L., Fitzgerald, D. A., Nathan, P. J., Moore, G. J., Uhde, T. W., & Tancer, M. E. (2005). Neural substrates for voluntary suppression of negative affect: A functional magnetic resonance imaging study. *Biological Psychiatry*, *57*, 210–219. https://doi.org/10.1016/j.biopsych.2004.10.030, PubMed: 15691521

Plassmann, H., O'doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, *27*, 9984–9988. https://doi.org/10.1523/JNEUROSCI.2131-07.2007, PubMed: 17855612

Platt, M. L., & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under uncertainty. *Nature Neuroscience*, *11*, 398–403. https://doi.org/10.1038/nn2062, PubMed: 18368046

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*, 59–63. https://doi.org/10.1016/j.tics.2005.12.004, PubMed: 16406760

Quay, H. C. (1997). Inhibition and attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology*, *25*, 7–13. https://doi.org/10.1023/A:1025799122529, PubMed: 9093895

Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, *38*, 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030, PubMed: 25938726

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338–342. https://doi.org/10.1006/ccog.1999.0386, PubMed: 10487787

Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, *1*, 563–581. https://doi.org/10.1007/s13164-010-0039-7, PubMed: 22558063

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology*, *96*, 538–558. https://doi.org/10.1037/a0014038, PubMed: 19254102

Riggall, A. C., & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *Journal of Neuroscience*, *32*, 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012, PubMed: 22993416

Rota, G., Sitaram, R., Veit, R., Erb, M., Weiskopf, N., Dogil, G., et al. (2009). Self-regulation of regional cortical activity using real-time fMRI: The right inferior frontal gyrus and linguistic processing. *Human Brain Mapping*, *30*, 1605–1614. https://doi.org/10.1002/hbm.20621, PubMed: 18661503

Samson, D., Apperly, I. A., & Humphreys, G. W. (2007). Error analyses reveal contrasting deficits in "theory of mind": Neuropsychological evidence from a 3-option false belief task. *Neuropsychologia*, *45*, 2561–2569. https://doi.org/10.1016/j.neuropsychologia.2007.03.013, PubMed: 17451756

Samson, D., Houthuys, S., & Humphreys, G. W. (2015). Self-perspective inhibition deficits cannot be explained by general executive control difficulties. *Cortex*, *70*, 189–201. https://doi.org/10.1016/j.cortex.2014.12.021, PubMed: 25752979

Saxe, R., & Baron-Cohen, S. (2006). The neuroscience of theory of mind. *Social Neuroscience*, *1*, 1–9. https://doi.org/10.1080/17470910601117463, PubMed: 18633783

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *Neuroimage*, *19*, 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1, PubMed: 12948738

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron*, *76*, 677–694. https://doi.org/10.1016/S1053-8119(03)00230-1, PubMed: 12948738

Schul, Y., & Burnstein, E. (1985). When discounting fails: Conditions under which individuals use discredited information in making a judgment. *Journal of Personality and Social Psychology*, *49*, 894–903. https://doi.org/10.1037/0022-3514.49.4.894

Sebastian, A., Pohl, M. F., Klöppel, S., Feige, B., Lange, T., Stahl, C., et al. (2013). Disentangling common and specific neural subprocesses of response inhibition. *Neuroimage*, *64*, 601–615. https://doi.org/10.1016/j.neuroimage.2012.09.020, PubMed: 22986077

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences, U.S.A.*, *93*, 628–633. https://doi.org/10.1073/pnas.93.2.628, PubMed: 8570606

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*, 1916–1936. https://doi.org/10.1152/jn.2001.86.4.1916, PubMed: 8570606

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*, 617–627. https://doi.org/10.1093/brain/awn279, PubMed: 18971202

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 17058–17062. https://doi.org/10.1073/pnas.1205828109, PubMed: 23011798

Shulman, G. L., Ollinger, J. M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Petersen, S. E., et al. (1999). Areas involved in encoding and applying directional expectations to moving objects. *Journal of Neuroscience*, *19*, 9480–9496. https://doi.org/10.1523/JNEUROSCI.19-21-09480.1999, PubMed: 10531451

Spinoza, B. (1982). *The ethics and selected letters* In S. Feldman (Ed.), & S. Shirley (Trans.). Indianapolis, IN: Hackett. (Original work published 1677).

Spitzer, M., Braun, U., Hermle, L., & Maier, S. (1993). Associative semantic network dysfunction in thought-disordered schizophrenic patients: Direct evidence from indirect semantic priming. *Biological Psychiatry*, *34*, 864–877. https://doi.org/10.1016/0006-3223(93)90054-H, PubMed: 8110913

Stawarczyk, D., & D'Argembeau, A. (2015). Neural correlates of personal goal processing during episodic future thinking and mind-wandering: An ALE meta-analysis. *Human Brain Mapping*, *36*, 2928–2947. https://doi.org/10.1002/hbm.22818, PubMed: 25931002

Street, C. N., & Kingstone, A. (2017). Aligning Spinoza with Descartes: An informed Cartesian account of the truth bias. *British Journal of Psychology*, *108*, 453–466. https://doi.org/10.1111/bjop.12210, PubMed: 27511287

Street, C. N., & Richardson, D. C. (2015). Descartes versus Spinoza: Truth, uncertainty, and bias. *Social Cognition*, *33*, 227–239. https://doi.org/10.1521/soco.2015.33.2.2

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*, 363–375. https://doi.org/10.1038/nrn1666, PubMed: 15832198

Swick, D., Ashley, V., & Turken, U. (2011). Are the neural correlates of stopping and not going identical? Quantitative meta-analysis of two response inhibition tasks. *Neuroimage*, *56*, 1655–1665. https://doi.org/10.1016/j.neuroimage.2011.02.070, PubMed: 21376819

Szczepanski, S. M., Pinsk, M. A., Douglas, M. M., Kastner, S., & Saalmann, Y. B. (2013). Functional and structural architecture of the human dorsal frontoparietal attention network. *Proceedings of the National Academy of Sciences, U.S.A.*, *110*, 15806–15811. https://doi.org/10.1073/pnas.1313903110, PubMed: 24019489

Thorson, E. (2015). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*, 460–480. https://doi.org/10.1080/10584609.2015.1102187

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*, 515–518. https://doi.org/10.1126/science.1134239, PubMed: 17255512

Tsujii, T., Masuda, S., Akiyama, T., & Watanabe, S. (2010). The role of inferior frontal cortex in belief-bias reasoning: An rTMS study. *Neuropsychologia*, *48*, 2005–2008. https://doi.org/10.1016/j.neuropsychologia.2010.03.021, PubMed: 20362600

Tsujii, T., Sakatani, K., Masuda, S., Akiyama, T., & Watanabe, S. (2011). Evaluating the roles of the inferior frontal gyrus and superior parietal lobule in deductive reasoning: An rTMS study. *Neuroimage*, *58*, 640–646. https://doi.org/10.1016/j.neuroimage.2011.06.076, PubMed: 21749923

Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al. (2001). I know what you are doing: A neurophysiological study. *Neuron*, *31*, 155–165. https://doi.org/10.1016/S0896-6273(01)00337-3, PubMed: 11498058

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 219–230. https://doi.org/10.1037/0278-7393.33.1.219, PubMed: 17201563

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38. https://doi.org/10.1016/j.concog.2008.09.006, PubMed: 18980847

Van Der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*, 460–467. https://doi.org/10.1038/s41591-022-01713-6, PubMed: 35273402

Vanni, S., Tanskanen, T., Seppä, M., Uutela, K., & Hari, R. (2001). Coinciding early activation of the human primary visual cortex and anteromedial cuneus. *Proceedings of the National Academy of Sciences, U.S.A.*, *98*, 2776–2780. https://doi.org/10.1073/pnas.041600898, PubMed: 11226316

van Oers, C. A., Vink, M., van Zandvoort, M. J., van der Worp, H. B., de Haan, E. H., Kappelle, L. J., et al. (2010). Contribution of the left and right inferior frontal gyrus in recovery from aphasia. A functional MRI study in stroke patients with preserved hemodynamic responsiveness. *Neuroimage*, *49*, 885–893. https://doi.org/10.1016/j.neuroimage.2009.08.057, PubMed: 19733673

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). Chichester: Wiley.

Windschitl, P. D., Scherer, A. M., Smith, A. R., & Rose, J. P. (2013). Why so confident? The influence of outcome desirability on selective exposure and likelihood judgment. *Organizational Behavior and Human Decision Processes*, *120*, 73–86. https://doi.org/10.1016/j.obhdp.2012.10.002

Windschitl, P. D., Smith, A. R., Rose, J. P., & Krizan, Z. (2010). The desirability bias in predictions: Going optimistic without leaving realism. *Organizational Behavior and Human Decision Processes*, *111*, 33–47. https://doi.org/10.1016/j.obhdp.2009.08.003

Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W. F., et al. (2005). Role of the contralateral inferior frontal gyrus in recovery of language function in poststroke aphasia: A combined repetitive transcranial magnetic stimulation and positron emission tomography study. *Stroke*, *36*, 1759–1763. https://doi.org/10.1161/01.STR.0000174487.81126.ef, PubMed: 16020770

Yacubian, J., Sommer, T., Schroeder, K., Gläscher, J., Kalisch, R., Leuenberger, B., et al. (2007). Gene–gene interaction associated with neural reward sensitivity. *Proceedings of the National Academy of Sciences, U.S.A.*, *104*, 8125–8130. https://doi.org/10.1073/pnas.0702029104, PubMed: 17483451