Believing, Desiring, or Just Thinking About: Toward a Neuroscientific Account of Propositional Attitudes

Believing, Desiring, or Just Thinking About:

Toward a Neuroscientific Account of Propositional Attitudes

A dissertation presented

by

Regan Marjorie Bernhard

to

The Department of Psychology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May 2017

Dissertation Advisor: Professor Joshua D. Greene **Regan Marjorie Bernhard**

Believing, Desiring, or Just Thinking About:
Toward a Neuroscientific Account of Propositional Attitudes

Abstract

Human minds can relate to a single idea in different ways. I can *believe* that Starbucks now sells donuts, but I can also *hope* or *fear* that Starbucks now sells donuts. Propositions, such as *Starbucks now sells donuts* are potential states of the world, or ideas thereof, that can be either true or false (McGrath, 2014). Propositional attitudes, on the other hand, are the mental states held by an agent toward a proposition (McKay & Nelson, 2014). The human mind has a remarkable capacity to hold a range of attitudes about an effectively infinite number of potential propositions. But how does our brain appropriately connect an attitude, such as believing or desiring, to a proposition? In the present research we use functional magnetic resonance imaging to identify brain regions that contain information about a limited set of object-location propositions (for example "the dog is in the upper right corner"), as well as regions associated with believing, desiring, or merely thinking about those propositions. We find increased neural activation associated with desiring a proposition in portions of the default mode network (left medial prefrontal cortex, posterior inferior parietal lobule, and middle temporal gyrus) as well as in the left dorsolateral prefrontal cortex. We find increased activation associated with believing a proposition in the right posterior parietal cortex and with thinking about a proposition that is not necessarily believed in the right inferior frontal gyrus. We find reliably decodable location information in the occipital lobe and object information in the right parahippocampal gyrus that is stable across propositional attitude. We also find that

that attitude has an effect on proposition representation. When a proposition is being desired, rather than merely being thought about, we find more object decodability in the left putamen, cingulate, cuneus, precuneus, and medial frontal gryus. When a proposition is being believed, rather than merely being thought about, we find more object decodability in the right posterior superior temporal gyrus, right paracentral lobule, and bilateral cuneus. Finally, we take the first steps towards developing a model for how the brain flexibly combines propositional attitudes and propositions by using functional connectivity to identify correlations between attitude-associated neural activation and attitude-specific object representation. We find a small but significant correlation between the activation in the left medial prefrontal cortex and the object representation in the left middle temporal gyrus when the object is a constituent of a desired proposition. We also find a small but significant correlation between activation in the right inferior frontal gyrus and object representation in the right posterior superior temporal gyrus when object is a constituent of a proposition that is being thought about but not necessarily believed.

TABLE OF CONTENTS

Acknowledgments

I would like to start by thanking my advisor, Josh Greene, for working with me on this project. I am especially grateful to have a mentor and collaborator who gets as wildly excited about potential discoveries as I do. Throughout this project, Josh has been continually patient, optimistic, and supportive. More generally, Josh has been an exceptionally kind, caring, and compassionate advisor and a very good friend.

I would also like to thank the members of my committee who have all played important roles both in the completion of this project and in my graduate school experience. I am grateful to Dan Gilbert for providing the thinking and writing that inspired large parts of this project. Liz Spelke has put tremendous thought into this project and has raised many challenging issues. Fiery Cushman's insight has been invaluable since the inception of this project, pushing me to be continually critical of my work. Further, over the past seven years, Fiery has been a ceaselessly supportive mentor and has provided me with a model of the kind of scientist I continue to work towards becoming.

Thanks to my lab brothers, Joe Paxton, Steven Frankland, Donal Cahill, Alek Chakroff, and Dillon Plunket, who have been a uniquely kind, thoughtful, and very sweet collection of friends. Other Greene lab members, past and present, Dave Rand, Elinor Amit, Ann Carroll, Morgan Henry, Sarah Gottlieb, and Arunima Sarin have also all played important roles in what has been a very special academic family. I would also like to thank Sarah Cotterill for her friendship and on-call statistics support.

I owe a tremendous debt of gratitude to my office mate, Steven Frankland. In addition to being an intellectual collaborator on this project, Steven painstakingly taught me how to do every piece of neuroimaging analysis included in this paper. There seemed to be no limit to his willingness to help me, and this project simply would not have happened without him. More importantly though, for the past seven years Steven has provided me with a steady, trustworthy, unconditional friendship that has made every day a little bit better.

I would like to thank my family, Alan, Kurt, Melanie, and Andrew Bernhard, Chelsi Peck, and Mike Enwall for being my biggest fan club, even when they have no idea I am doing. I also owe special thanks to my mom, Suzanne Pinto, for listening to me talk about my work and asking me hard questions, and for relentlessly seeing me as awesome, even when I am mired in my own doubts.

Finally, I owe my deepest thanks to my son, Finny Bernhard-Krol, and my wife Cindy Krol. To Finny, for giving up so much of his time with me so that I could write, or program, or worry. And to Cindy, for standing by me when I left a good paying and secure job to be a volunteer research assistant with the hope of one day getting into graduate school. For holding our lives together when I have been too busy working or thinking to remember to do it myself. For taking a deep breath and saying ok every time I said, "I think it's just going to be one more year." And most of all, for her love, friendship, and for giving me something much more important to come home to.

I. INTRODUCTION AND LITERATURE REVIEW

Introduction

Humans are able to have many different thoughts or attitudes about a single idea. For example, you may *know* that Starbucks now sells donuts. You could also *hope* that Starbucks now sells donuts, or *doubt* that Starbucks now sells donuts, or even just *think about* the possibility that Starbucks now sells donuts. Propositions, such as *Starbucks now sells donuts*, are states of the world, or sharable ideas about states of the world, that can be either true or false. Propositions can range from the concrete and contingently true (my dog has four legs), to the abstract and necessarily true (2+2=4), and can be certainly knowable (this paper was typed using a computer) or possibly unknowable (there is life after death). Further, propositions can be limitlessly combined to form an infinite number of more complex propositions (my four-legged dog typed this paper using a computer; (McGrath, 2014).[1]

Propositional attitudes, on the other hand, are the mental states held by an agent toward a proposition (McKay & Nelson, 2014). Propositions can be believed, feared, wanted, hoped for, etc. The human mind has the remarkable capacity to hold a full range of attitudes[2] about any of an effectively infinite number of potential propositions. But what is the relationship between propositions and the attitudes we hold about them in the mind, and ultimately the brain? The present work aims to better understand how propositions and propositional attitudes are flexibly combined in the brain. More specifically, we seek to understand (1) how propositional information is represented such

[1] For the present work, we consider propositions to be the *meaning* referred to by propositional *that*-clauses, not the clauses themselves.

[2] Throughout this paper use the term *attitude* as shorthand for propositional attitude. We are not referring to attitudes in the more colloquial, social psychological sense, where attitudes are valenced evaluations of specific objects, people, events, ideas, etc. (Allport, 1935).

that it may be combined with different propositional attitudes (2), how the mental acts of believing, desiring, and merely thinking about are instantiated in the brain such that they may be combined with different propositions, and (3) whether or how believing, desiring, or merely thinking about affects the representation of propositions—the objects of our beliefs, desires, and thoughts.

Theoretical Background

According to a commonsense view of the mind, sometimes called "folk psychology," people's propositional attitudes (what they believe, desire, fear, etc.) are the unseen causes of ordinary behavior. Philosophers such as Dretske (1988) and Fodor (1987) argue that this commonsense view of the mind provides a rough, but fundamentally accurate, guide to how thoughts produce goal-directed behavior. They propose that we have no other way of making sense of people's behavior than by making ascriptions to such propositional attitudes. Within this framework, we can then develop more elaborate, and possibly accurate, theories about the relationships between propositional attitudes and behavior (Sellars, 1963). The indispensability of attitude attributions when making sense of behavior leads to the conclusion that propositional attitudes must denote some underlying mental state. As Fodor says in his 1987, *Psychosemantics*,

> It is a deep fact about the world that the most powerful etiological generalizations hold of unobservable causes. It is thus a test of the depth of a theory that many of its generalizations subsume interactions among unobservables. Commonsense psychology passes this test. It takes for granted that overt behavior comes at the end of a causal chain whose links are mental events – hence unobservable – and which may be arbitrarily long.

This is a central claim of the Representational Theory of Mind. Belief, desire, fear, etc. involve someone's having in her mind, and ultimately her brain, belief-like, desire-like, or fear-like "relations" to a representation with the same propositional content as the belief, desire, fear, etc. (Fodor, 1981, 1987; Schiffer, 1981; Sterelny, 1990). Specifically, Fodor's *Language of Thought Hypothesis* claims that for each propositional attitude, there is a unique and distinct relation between the individual who holds that attitude and the mental representation of a proposition (Aydede, 2015; Fodor, 1987). For example, I believe Starbucks now sells donuts, if, and only if, there is a belief relation between me and a mental representation, the content of which is *Starbucks now sells donuts*. Fodor also argues that the proposition representations are language-like in structure in that they have a combinatorial syntax and semantics (Egan, 1991; Fodor, 1987). That is, representations have content that corresponds to something in the world and these representations can be combined to form increasingly complex beliefs, desires, etc.

Some connectionist theories provide an alternative to classical representational approaches such as the Language of Thought Hypothesis. In general, connectionist theories make three main claims: 1) information is encoded in connection weights between widely distributed hidden units, 2) the individual units in the network have no symbolic interpretation, and 3) learning algorithms adjust the connection weights to make the network function more effectively (Ramsey, Stich, & Garan, 1990; Rumelhart, McClelland, & Group, 1988; Smolensky, 1988).

Many tenets of connectionism are uncontroversial. Essentially all contemporary scientists and philosophers accept that the mind is realized by the brain, that the brain is a

complex neural network, and that the brain's function is adjusted through the modification of connection weights among neurons. Likewise, essentially all agree that mental representations are to some extent distributed across nodes (i.e., neurons) in the brain. Where some connectionist theories diverge from those consistent with folk psychology is in claiming that there is no level at which the brain implements propositional attitudes and propositions as functionally distinct elements within an explicitly compositional framework (Smolensky, 1988).

The most forceful opposition to representational approaches to propositional attitudes comes form Eliminative Materialism (EM). Strong eliminative materialists believe that the terms used to refer to propositional attitudes fail to denote anything that actually exists (Churchland, 1981; Ramsey et al., 1990). As described by Churchland (1981),

> Eliminative materialism is the thesis that our common sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience.

Supporters of EM argue that folk psychology is an empirical theory, and as such, can be subjected to the same degree of scrutiny as other empirical theories. Therefore, like all other empirical theories, folk psychology can be proven to be false (Churchland, 1981; Lycan & Pappas, 1972).

Churchland provides several arguments to support his claim that folk psychology is a "false theory." First, he proposes that we can understand others' behaviors, not by making attributions to propositional attitudes such as beliefs or desires, but by calling on a body of "lore" that comprises folk psychology. In other words, we expect particular

behaviors under specific conditions because we have culturally derived knowledge that those conditions usually lead to those behaviors (this view is also put forth by Sellars, 1956, in his early work). Continued experience engaging with others grows this corpus of knowledge and improves our ability to predict behavior in the future. Churchland also argues that simply because it feels as if we have beliefs, desires, etc. does not require that others do as well. What's more, he asserts that our understandings of the workings of our own minds are imperfect at best, making them an unreliable indicator of what might be driving others' behaviors.

Proponents of EM argue that neuroscience will ultimately prove that mental states cannot be the product of neurological processes because nothing beyond these neurological processes exists (Churchland, 1981; Cornman, 1968; Feyerabend, 1963; Rorty, 1965). Yet eliminative materialists usually fall short of offering a proposal for what these physical processes encompass at a cognitive level. In his earlier work, Stich and colleagues (Ramsey et al., 1990) argued that connectionist models, which entail the highly distributed storage of information, might ultimately undermine representationalist approaches like the Language of Thought Hypothesis, which tend to describe representation as more modular. However, in his later work Stich concedes that while connectionist models offer an alternative to theories of mental representation that embrace folk psychology, they do not render propositional attitudes non-existent. Instead they offer alternative models of how these mental states are instantiated in the brain, failing to meet EM's standard of proving folk psychology to be radically false (Stich, 1999).

The theoretical motivation for the present research is consistent with the folk psychology view of propositional attitudes. As a starting point, we hypothesize that propositional attitudes denote mental states that are realized as distinct processes in the brain. Importantly, we are not asserting that belief or desire are mental states that are necessarily different from those that are commonly associated with belief or desire *related* processes. For example, desire need not be some special mental state distinct from those related to reward anticipation, affect, valuation etc. As such we might expect to see that the neural activation associated with desire is the very activation we see associated with reward anticipation, affect, valuation, etc. Rather, in keeping with a folk psychological perspective, we suggest that the neural processes associated with each propositional attitude can be combined in a flexible way with different propositions. Likewise, in keeping with the Representational Theory of Mind, we suppose that information about states of the world, which some would call "propositions," is represented in the brain in such a way that it can be flexibly combined with different propositional attitudes. In this project we use functional neuroimaging to better understand how propositional attitudes are realized in the brain, and how propositional content may be represented in such a way that allows us to hold different attitudes about the very same proposition.

Why Belief and Desire?

In her work on intention, Anscombe (1957) draws a distinction between two ways words can relate to the world; those that aim to match the state of the world ("word-to-world"), and those that aim to alter the world in some way to match our words ("world-to-word"). Often called *direction-of-fit,* this relationship between words and the world has also been applied to understanding mental states, distinguishing those with a "mind-to-world" direction of fit from those with a "world-to-mind" direction of fit (Searle, 1983).

Within this framework, belief and desire serve as canonical examples of these two categories of mental states. Beliefs, with their mind-to-world direction of fit, function to represent true information about the world. Therefore, beliefs should change to match the world. On the other hand, desires, with their world-to-mind direction of fit, function to realize a world that fits them. In other words, desires drive us to change the world to match our mental states (Humberstone, 1992; Searle, 1983; Smith, 1987). Yet despite their opposing directions-of-fit, we are able to hold these distinct attitudes[3] about the very same states of the world. How then do attitudes that serve different functions in our interactions with the world share propositional content? We take some first steps towards answering this question by examining the neural processes associated with believing,

[3] Alternatively, the *desires-as-belief* view argues that some types of beliefs can be motivational and that some, if not all desires might be a kind of belief. Here, desires are not a distinct attitude from belief, but rather beliefs about "what would be good" (Price, 1989). Given that traditional direction-of-fit views argue that mind-to-world and world-to-mind form mutually exclusive categories of attitude classification (Smith, 1987), the blurring of beliefs and desires renders the direction-of-fit hypothesis untenable. However, strong criticism of the desires-as-belief view comes from Lewis (1988) and Collins (1988) who identify several ways in which it is incompatible with Bayesian decision theory and belief revision based on counterfactual reasoning.

desiring, and merely thinking about the same set of propositions. First, we investigate how belief and desire are instantiated in the brain. Then, we examine the effect of each attitude on proposition representation with the aim of better understanding how the brain flexibly combines propositional content with propositional attitudes.

Belief and Desire in the Brain

Desire

The extensive body of research on the neural processing of reward gives us some insight into how desire is instantiated in the brain. Some of the earliest research characterizing reward systems in the brain came from Olds and Milner (1954). They discovered that rats would repeatedly perform the same behavior in order to self-administer brief bursts of electrical stimulation to the septal area (part of the limbic system). Subsequent research using rodent models isolated a network of structures associated with reward, identifying dopaminergic neurons that project from the midbrain to the nucleus accumbens in the ventral striatum (Koob, 1992; Pfaus et al., 1990). Primate research and early neuroimaging studies in humans expanded the reward network beyond the midbrain and ventral striatum to include the amygdala (Nishijo, Ono, & Nishino, 1988), orbitofrontal, and medial prefrontal cortices (Thorpe, Rolls, & Maddison, 1983; Thut et al., 1997; Tremblay & Schultz, 2000; Watanabe, 1996).

Much of the reward-related research over the intervening thirty years has been focused on disentangling the specific functionality of discrete regions within this network. Single unit recordings in non-human primates have found that neurons in the orbital frontal cortex (OFC) track with reward magnitude (Critchley & Rolls, 1996;

O'Doherty et al., 2000). These findings have been confirmed by neuroimaging studies in humans, which find that that activation in the OFC tracks with the value of monetary (Elliott, Newman, Longe, & Deakin, 2003) and social rewards (Lin, Adolphs, & Rangel, 2012), as well as with taste (O'Doherty, Rolls, Francis, Bowtell, & McGlone, 2001; Small et al., 2003a), olfactory (Anderson et al., 2003; Gottfried, O'Doherty, & Dolan, 2002), and auditory rewards (Blood, Zatorre, Bermudez, & Evans, 1999). The amygdala has also been implicated in processing reward value. This region responds to both aversive and pleasant stimuli (Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002; Holland & Gallagher, 2004; Morris, Frith, Perrett, & Rowland, 1996; Zald & Pardo, 1997) suggesting it codes affective value in particular. However, work intended to differentiate the neural response to reward valance versus reward intensity found that activation in the OFC was associated with reward valance, while the amygdala responded specifically to reward intensity (Anderson & Sobel, 2003; Small et al., 2003).

The OFC and other regions of the ventromedial prefrontal cortex (vmPFC) play a role in encoding the value of potentially rewarding stimuli, supporting reward motivated decision-making. Neuroimaging studies in humans have found that these regions track with the predictive value of stimuli during decision-making (Kable & Glimcher, 2007) including in economic transactions (Chib, Rangel, Shimojo, & O'Doherty, 2009; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Plassmann, O'Doherty, & Rangel, 2007), and under risk (Levy, Snell, Nelson, Rustichini, & Glimcher, 2010; Tom, Fox, Trepel, & Poldrack, 2007).

Early work in primates by Schultz, Dayan, and Montague (1997) suggests that the ventral striatum is not tracking the predictive value of the stimuli itself or the actual value

of the ultimate reward, but rather the discrepancy between the predicted and actual reward (i.e. the "prediction error"). In this study, monkeys were trained to respond to a conditioned stimulus for a reward. Schulz and colleagues found that after conditioning, dopaminergic neurons in the striatum responded not to the reward, but to the stimulus. Further if an expected reward was not received, firing in these neurons dropped below baseline. A large body of neuroimaging research in humans supports the idea that the ventral striatum specifically encodes prediction error signals (Berns, McClure, Pagnoni, & Montague, 2001; Delgado, Nystrom, Fissell, Noll, & Fiez, 2000; Hare et al., 2008; J. P. O'Doherty, 2004; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Seymour, Daw, Dayan, Singer, & Dolan, 2007; Yacubian et al., 2006).

Most recently Berridge and Robinson have differentiated between two separable components of reward: "liking", or the hedonic response to rewarding stimuli, and "wanting", which they describe as "a type of incentive motivation that promotes approach toward and consumption of rewards" (Berridge, Robinson, & Aldridge, 2009). They argue that under most circumstances, rewards are usually liked and wanted to the same degree, making them part of the same process. However, under some circumstances (for example with addictive behavior) rewards may be wanted without necessarily being liked. Berridge and colleagues find that neural activation associated with liking is restricted to the nucleus accumbens and ventral pallidum (Berridge, 2009; Berridge et al., 2009) and does not extend into the rest of the reward network.

In the task used in the present study, participants receive a monetary reward if a "target object" moves to a specific, "target location" on the screen. However, they know there is only a 25% chance that target object will end up in the target location. We are

particularly interested in the brain activation associated with the period of time when subjects are actively desiring (or hoping) that the target object has gone to the target location, but before they learn the actual outcome. Desire, as elicited by our task, may involve several different reward-related processes. First, because desire involves the anticipation of a potentially rewarding outcome rather than receipt of reward, we might expect to see neural activation in regions associated with reward anticipation, rather than reward receipt. A meta-analysis of 12 studies finds that reward anticipation, in particular, is associated with increased activation in the right anterior cingulate, nucleus accumbens, insula, caudate, supplementary motor area, thalamus, and culmen (Knutson & Greer, 2008). We might also expect to see neural activation in regions associated with stimulus valuation such as the OFC and other regions of the vmPFC (Berridge, 2009; Chib et al., 2009; Hare et al., 2008; Kable & Glimcher, 2007; Levy et al., 2010; Plassmann et al., 2007; Tom et al., 2007). However, subjects in our task are anticipating a reward under uncertainty. Non-human primate research has found that clusters of dopaminergic neurons in the ventral striatum increase their firing rate as the probability of receiving an expected reward decreases (Fiorillo, Tobler, & Schultz, 2003) and neuroimaging studies in humans have found that the midbrain, ventral striatum (Abler, Walter, Erk, Kammerer, & Spitzer, 2006; Yacubian et al., 2007), and medial prefrontal cortex can all be responsive to reward uncertainty (Critchley, Mathias, & Dolan, 2001; Dreher, Kohn, & Berman, 2006; O'Neill & Schultz, 2010).

Belief

Philosophers use the term "belief" to refer to the attitude we have when we regard something as true (Schwitzgebel, 2015). Belief is often divided into two categories. Occurent beliefs are those that are actively being entertained. On the other hand, dispositional beliefs are those that an individual holds, but that are not actively being thought about. For example, I may hold the belief that horses are measured in hands, but only rarely (vary rarely) does this belief come to the forefront of my mind. When it does, it is an occurent belief. The rest of the time, I possess it only dispositionally (Audi, 1994).

Some philosophers also differentiate between implicit and explicit beliefs (Schwitzgebel, 2015). Explicit beliefs are those that we have consciously considered. From a representationalist perspective, they are those propositions that have been represented in a belief-like manner. Implicit beliefs, on the other hand, are those that we possess, but have not explicitly considered. Dennett (1978) suggests that implicit beliefs may simply be beliefs that are quickly derivable from explicit beliefs. If I explicitly believe the population of Cambridge, MA is around 105,000, I also hold the implicit belief that the population of Cambridge is less then 200,000, even if I have never explicitly had that particular thought. The implicit belief does not initially require a formal representation because I can quickly derive it from my explicit belief. Fodor (1987) posits a similar relationship between explicit and implicit beliefs, suggesting that at least some implicit beliefs may emerge from more basic structural facts. For example, I may hold the explicit beliefs that the team with the most points wins a basketball game and points are earned for your team by having a player on your team shoot the ball through the basket. From these two beliefs a further belief may emerge – for example the

belief that I should not let someone on the other team take the ball away from me – even if I do not explicitly represent that particular belief.

In this project, we are particularly interested in how explicit, occurent beliefs about propositions are manifest in the brain. While a broad spectrum of research in cognitive neuroscience examines processes involved in belief formation and retention (e.g. perception, learning, memory, conceptual representation, etc.), there is relatively little research on belief per se – that is, on understanding how believing a proposition differs from, and is similar to, having other propositional attitudes toward that same proposition. Gilbert (1991) puts forth two competing theories on the relationship between proposition representation and belief. According to what he calls the Cartesian theory, propositions that are believed must first be understood and represented, and only then are they assessed for veracity and, if deemed true, believed. Alternatively, the Spinozan theory posits that to understand and represent a proposition is to believe it. It is only disbelief that requires an additional processing step beyond comprehension.

In a series of experiments, Gilbert and colleagues provide support for the Spinozan view by demonstrating that putting individuals under cognitive load or time pressure increases the likelihood of believing information presented as false (Gilbert, Tafarodi, & Malone, 1993). They argue that their interventions prevented subjects from taking the secondary step of unbelieving false information that has been automatically believed. Additional support for the Spinozan view comes from research on the illusory truth effect (Begg, Anas, & Farinacci, 1992; Dechêne, Stahl, Hansen, & Wänke, 2010) in which individuals are more likely to report a statement as true if they have been exposed to it before. Moreover, the magnitude of this effect is decreased when individuals engage

more critically with the information during encoding than when they process the information more superficially (Hawkins & Hoch, 1992).

If the Spinozan theory is correct, as the behavioral evidence suggests, a proposition that is believed should be associated with no neural activation beyond the mere representation of the proposition (see the following section for a discussion on the neural instantiation of proposition representation). However, when propositions are not believed (either because they are disbelieved or because it is unknown whether or not they are true) we should see additional neural activation associated with "unbelieving" the initial automatic belief. The behavioral evidence described above suggests that this process of unbelieving would likely involve brain regions associated with cognitive control such as the dorsolateral prefrontal cortex, frontopolar cortex, inferior frontal gyrus, and anterior cingulate cortex (Aron, Behrens, Smith, Frank, & Poldrack, 2007; Chein & Schneider, 2005; Cole & Schneider, 2007; Niendam et al., 2012). In fact, findings from neuroimaging research, in which participants are asked to evaluate the truth content of statements, seem to support the Spinozan theory. One such study finds that processing false, but not true statements, increases activation in the frontopolar cortex (Marques, Canessa, & Cappa, 2009). A second set of studies found reduced activation in the vmPFC when evaluating statements as false (Harris et al., 2009; Harris, Sheth, & Cohen, 2008). The vmPFC is a key part of the default mode network, which is deactivated when engaging in cognitively taxing or attention demanding tasks (Gusnard, Akbudak, Shulman, & Raichle, 2001; Mazoyer et al., 2001; Shulman et al., 1997). Harris and colleagues argue, therefore, that the reduced activation in the vmPFC when evaluating false statements is indicative of cognitive control engagement. Alternatively, if

the Cartesian theory is correct, we would expect believed propositions to be associated with some neural marker identifying them as distinct from propositions that are merely being thought about.

A key feature of the set of neuroimaging studies described above is that they ask subjects to consciously report or reflect on their own beliefs. Therefore, it is unclear whether the described patterns of neural activation reflect the beliefs themselves, or the metacognitive processes involved in assessing one's own beliefs. In the present project, we are investigating first-order beliefs, not the second-order assessment of the truth of a proposition or the strength of the belief. Only one study (Goel & Dolan, 2004) has attempted to look at the neural correlates of belief versus non-belief without having participants explicitly reflect on their beliefs. In this experiment, participants assessed logical arguments where the conclusions were either consistent with the participants' beliefs (for example, "all apples are red; all red foods are fruit; *all apples are fruit*") or were belief neutral. Belief-neutral syllogisms contained sentences that subjects could not have beliefs about because one or more words were unknown to the subjects (for example, "no codes are highly complex; some quipu are highly complex; no quipu are codes"). They found that subjects tended to endorse invalid syllogisms when the conclusion was consistent with their beliefs about the world (All calculators are machines; All computers are calculators; Some machines are not computers). They also found that engaging with the belief-laden logical arguments increased neural activation in the left pole of the middle temporal gyrus when compared to the belief-neutral arguments. Further, they identified increased activation in the right inferior frontal gyrus associated with performing well on the belief suppression portion of the task (recognizing

invalid syllogisms, even when the conclusion was consistent with their beliefs about the world). However, in this study, belief is confounded with propositional content. Belief-laden syllogisms were also the only completely comprehensible syllogisms, whereas belief-neutral syllogisms also contained unknown words. In fact, in each of the studies described above, separate sets of propositions are believed versus not, leaving open the possibility that neural activation thought to be associated with belief is actually a product of the propositions themselves. To our knowledge, there are no studies that examine the neural mechanisms of belief by systematically manipulating whether or not the same propositions are believed.

Believed and Desired Propositions in the Brain

As discussed above, many philosophical proponents of the Representational Theory of Mind (most notably Fodor, 1975, 1987) suggest that a proposition, *P*, is not believed unless it plays a belief-like role in cognition. Similarly, *P* is not desired unless it plays a distinct desire-like role. One popular instantiation of this theory posits that the mind contains a figurative "belief box" and "desire box" (Schiffer, 1981). To believe *P* requires that *P* is tokened in the "belief box". Likewise, to desire *P* requires that *P* is tokened in the "desire box." The strictest version of this hypothesis requires that *P* be a linguistic representation (Fodor, 1987) and/or that the actual symbolic token that means *P* is consistent as it is moved from box to box (Schiffer, 1981). A more liberal interpretation of this stance posits that the representation of a single proposition somehow changes depending on whether it is believed or desired. For the purposes of this project, we call this the *variable representation hypothesis*. In the brain, this might mean that

propositions are represented in different brain regions depending on whether they are believed or desired (a more literal interpretation of the "belief box" theory) or that the attitude has an effect on the nature of the propositional representation.

However, we would also expect some stable proposition representation despite the attitude being held about the proposition. In fact, one might argue that the relevant propositional representation must exist in some stable form. If not, it would not be true that it is the same proposition that is, for example, believed on one occasion but desired on another. Perhaps a good analogy for this *stable representation hypothesis* is amodal conceptual representation. While the representation of a single concept, for example "dog," can be specifically associated with the perceptual experience during encoding (Barsalou, Simmons, Barbey, & Wilson, 2003; Kiefer, 2005; Kiefer, Sim, Herrnberger, Grothe, & Hoenig, 2008; Kiefer, Sim, Liebich, Hauk, & Tanaka, 2007), there also appears to be patterns of brain activation common to the concept across input modalities (Fairhall & Caramazza, 2013; Simanova, Hagoort, Oostenveld, & Van Gerven, 2014). Likewise, if we do find that the representation of a proposition is moderated by attitude, we would almost certainly expect to also see some common representation for a single proposition despite attitude.

In the present project, subjects are induced to believe, desire, or think about a target object (a dog, mop, snake, or hose) being in a target location (upper left, upper right, bottom left, bottom right) on the screen. These object-location combinations comprise the propositions of interest in our study (for example *dog in upper right corner*). Where in the brain might we see this object-location proposition information?

One factor that may influence the proposition representation in our study is the level at which subjects encode the object and location information. Much of the information in the task is presented visually – in each trial subjects see pictures of the four objects and visual markers indicating the four locations. Therefore, at the lowest level of encoding we may find some representation of both object and location in the primary visual cortex (Kolb, Whishaw, & Teskey, 2014). Visual processing of objects extends along the occipototemporal cortical pathway through the posterior ventral temporal lobe. Visual processing related to motion and spatial information extends through the medial superior temporal lobes and into the inferior parietal lobes (Ungerleider & Haxby, 1994).

Likewise, the proposition information may be represented at higher levels within the extended visual system. Neuropsychological studies have linked damage to the left posterior temporal cortex to a loss of conceptual knowledge (Hart & Gordon, 1990; Mummery et al., 1999; Sharp, Scott, & Wise, 2004). TMS and neuroimaging studies have linked the fusiform gyrus and inferior and middle temporal gyri to the conceptual processing of words and images (Davis & Johnsrude, 2003; Rodd, Davis, & Johnsrude, 2005). The fusiform gyrus is especially important for representing categories of objects. Dissociable regions of the fusiform are more active for different categories of objects (for example, faces, houses, animals – see Grill-Spector, Kourtzi, and Kanwisher (2001) for a review), and patterns of neural activation further differentiate between a large number of categories (Cox & Savoy, 2003; Haxby et al., 2001; Spiridon & Kanwisher, 2002).

Given that the task used in the present project requires remembering objects and locations for several seconds, we may also see proposition-related activation in brain

regions associated with working memory. Most crucial for working memory are the prefrontal cortex, medial and inferior temporal cortices, and parts of the parietal cortex (Curtis & D'Esposito, 2003; Nystrom et al., 2000; Postle, Stern, Rosen, & Corkin, 2000; Ranganath & D'Esposito, 2005; Wager & Smith, 2003). Portions of the lingual and fusiform gyri, and inferior temporal cortex are preferentially activated for object working memory, while spatial working memory is associated specifically with activation in the superior parietal cortex (Wager & Smith, 2003).

The medial temporal lobe, and hippocampus and parahippocampal gyrus in particular, play a critical role in processing, binding, or storing location and object information. Neural recording studies in both humans and macaque monkeys have found that hippocampal cells fire in patterns that encode conjunctions between objects and their locations (Ekstrom et al., 2003; Rolls & Xiang, 2005). Neuroimaging studies in humans have also found greater hippocampal activation in tasks that require encoding, retrieving, or storing object locations in memory for a short period of time (Maguire et al., 1998; K. J. Mitchell, Johnson, Raye, & D'Esposito, 2000; Owen, Doyon, Petrides, & Evans, 1996; Piekema, Kessels, Rijpkema, & Fernández, 2009). Further, patients with bilateral hippocampal lesions cannot maintain object-location associations in memory for longer than a few seconds (Olson, Page, Moore, Chatterjee, & Verfaellie, 2006).

We want to reiterate here that when we use the term "proposition representation" we are not claiming that these representations necessarily qualify as "propositional" according to all definitions of this contentious term (e.g. Pylyshyn, 1973). Rather, we are looking to identify some representation of information about states of the world that can be combined with different attitudes in a structured, flexible way. Given this standard

– that the neural activation must represent information about which we can have beliefs or desires – one might argue that identifying activation associated with the low level sensory representation of the target object or location (for example patterns of neural activation that identify the shape of the target object in V1) is insufficient. Our goal then is to identify activation associated with the target object and location that suggests a deeper level of encoding, more conceptual than perceptual.

In this project we use functional neuroimaging to study the relationship between proposition representation and propositional attitudes in the brain. We begin by describing the shell game task used in this study. Then we investigate the neural activation associated specifically with believing, desiring, or thinking about a set of propositions. Next we look at patterns of neural activity associated with proposition representation, both across and within propositional attitudes. This allows us to look for evidence supporting both the *stable* and *variable representation hypotheses.* Finally, we take the first steps towards developing a model of how the brain integrates propositional attitudes and propositions.

II. THE SHELL GAME TASK AND BEHAVIORAL RESULTS

The Shell Game Task

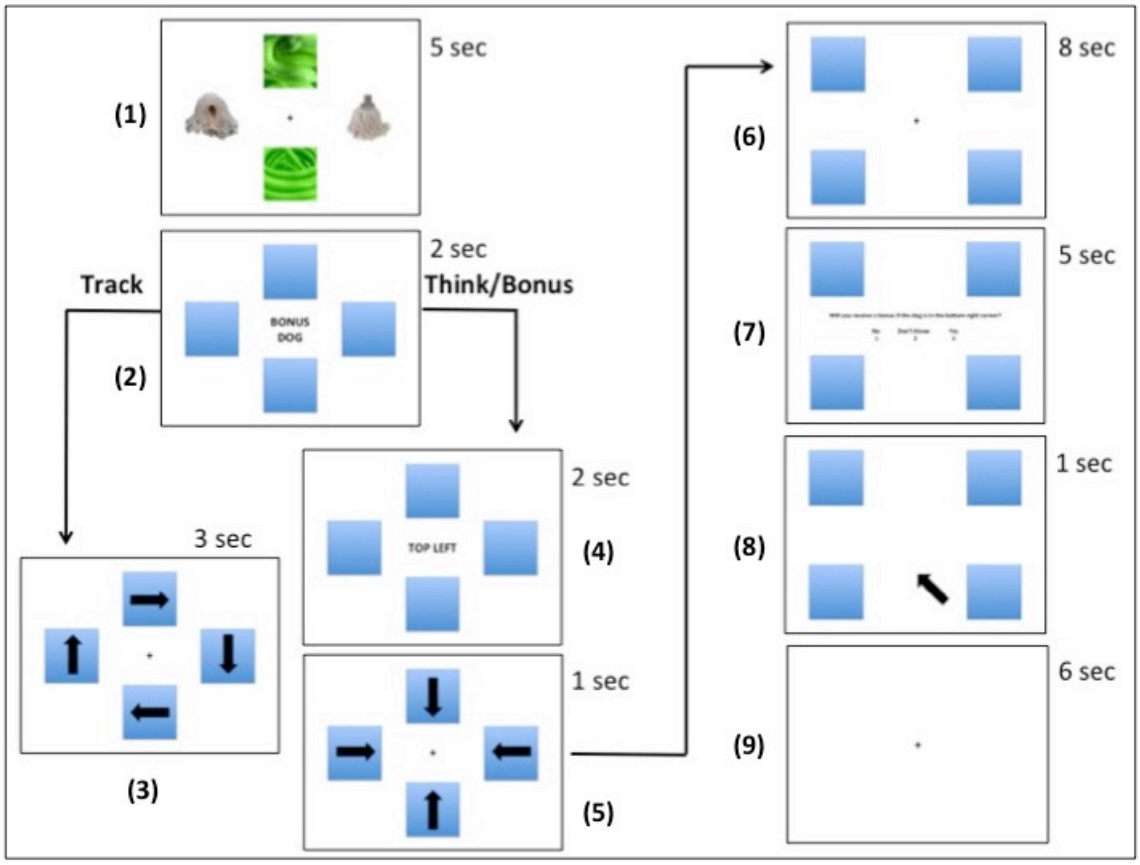


Figure 1. Task sequence and timing. **(1)** The objects are randomly placed in one of four locations (left, right, top, bottom) for 5 seconds. During this time subjects must learn the starting location of each object. **(2)** The objects are covered (for 2 seconds), and subjects learn the trial type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

For this study, we developed a novel visual object-tracking paradigm modeled after a classic shell game (see Figure 1 above). The shell game task is designed to experimentally manipulate the propositional contents of people's thoughts as well as the

attitudes adopted towards those propositions, such that a given proposition may be either believed to be true, desired, or merely thought about. By having subjects complete the shell game task while undergoing fMRI, we can observe patterns of neural activation associated with believing, desiring, and thinking about a set of propositions, with propositions randomly paired with each of the three propositional attitudes.

In the shell game task, pictures of four objects (a dog, a mop, a snake, and a hose) are presented on the screen. After five seconds, blue squares cover each object, and the trial type (described below) and the name of a "target object" are verbally presented. Finally, the objects are "shuffled" and distributed to each of the four corners of the screen.

There are three types of trials:

In *track* trials, the objects are shuffled in such a way that the target object is very easy to follow and its final location should be clear to the participant. In this condition, subjects should have a clear belief about the target object's final location, but no strong desire for it to be there.

In *bonus* trials, prior to the objects being shuffled, the subject is verbally informed about a "target location." Subjects know that if the target object ends up in this target location, they will receive a $5 bonus. The objects are then shuffled in such a way that it is impossible to track the target object. In this condition, subjects should have a strong desire for the target object to be in the target location, but have no particular belief about where the target object actually is.

The final trial type, the *think* trial, closely resembles the bonus trial in structure and visual content. After finding out the trial type and target object, subjects are given a target location and are told to merely think about the target object being in the target location. They key difference is that in think trials, subjects do not receive a bonus if the target object ends up in the target location. In this case, subjects neither know the target object's final location, nor have a strong reason to want it to be there.

Because the belief shuffles are always trackable, and the desire and think shuffles are not, shuffle trackability is confounded with attitude type. To mitigate the effects of this confound, we generated two different shuffle styles (described in detail below), one in which the objects rotate around the screen and another in which arrows point in the direction the objects move. Creating two shuffle types allows us to assess whether this confound, a product of incidental differences in the visual stimuli, could plausibly explain any results of interest. All of our key analyses were done collapsing across shuffle types.

In the *rotation* version of the shuffle, the blue squares rotate around the screen to arrive in their final locations in each of the four corners. For the track trials the squares move from their starting locations directly to their destination corners. In the bonus and think trials the squares converge in the center of the screen, completely overlapping before moving out to their destination corners. The convergence in the center makes it impossible to follow which square moves to which corner. Although the movement path is different for the track and desire/think trials, the rotation speed and total distance moved was matched as closely as possible between trial types. In the *arrow* version of the shuffle, an arrow appears over each blue square indicating the path the subject is supposed to imagine the square taking. In track trials, the arrows point to the squares'

destination corners. In the bonus and think trials the arrows point to the center of the screen, revealing no information about the squares' final destinations. In all three trial types, the arrows are displayed, then the squares disappear and reappear in their final destination. The shuffle version was randomly assigned to each trial, so that there were approximately equal numbers of trials with rotation and arrow shuffles in each attitude condition.

After the objects are shuffled, participants view a fixation/delay screen for eight seconds with the four blue squares in each corner and a fixation cross in the center. This is the time period we will be using in all of our key analyses. Critically, during this time subjects are viewing the same screen regardless of condition.

The delay screen is followed by one of the following attention check questions:

Is the [target object] in [some location]?

Will you get a bonus if the [target object] is in [some location]?

Were you asked to think about the [target object] being in [some location]?

Subjects 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 subjects only receive the question specifically probing the attitude 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. Further, while 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, subjects are required on all trials to attend to and remember the trial type, the target object, and the target location.

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 location. In track trials, the arrow tells subjects whether they correctly tracked the object. In bonus trials, it tells subjects whether they will be receiving a bonus on that trial. In the think trials the arrow merely informs the subject of the target object's final location.

Importantly, subjects never actually see any of the objects in their final locations. The objects begin each trial centered on each side of the screen but are covered before they are moved to their final locations in the four corners. Thus, this paradigm enables us to examine neural activity associated with beliefs, desires, or thoughts about hidden states that have not been directly perceived.

The experiment has 12 runs of 12 trials each. Each object serves as the "target object" three times during a run, and there are four repetitions of each trial type per run. Therefore, there are 48 appearances of each trial type and 36 appearances of each object as the target object across the entire experiment. Within each run, the four appearances of each trial type and three appearances of each target object are in random order. There are also 36 total appearances of each target location presented in random order throughout the entire experiment (as opposed to within run). Subjects receive a bonus in only 25% of the bonus trials (12/48) to minimize their ability to anticipate the final location of the target object in these trials. To this end, these 12 "win" trials are randomly distributed

throughout the entire experiment. A single trial takes 30 seconds, with a 30-second break between runs, making the duration of the entire experiment approximately 80 minutes.

Participants

Forty participants (24 female; age 19-64) were recruited through the Harvard Study Pool. All were native English speakers, right handed, and had no history of neurological problems. Of these, three were excluded for poor performance (less than 70% accuracy) on the attention check task. Six were excluded for excessive motion during scanning (see details in Appendix) and one was excluded because of technical issues leading to incomplete data. The remaining thirty subjects were included in both the behavioral and neuroimaging analyses described throughout the paper.

Behavioral Results

Subjects performed very well on the attention check questions, getting 94% correct overall. However, they did perform significantly better on questions following bonus and think trials than on those following track trials ($F(2,4311) = 3.23$, $p = .047$). Nevertheless, response accuracy following track trials was still high, at 93.1% (compared to 95.8% for desire and 94.9% for think trials).

One potential concern about this task is that subjects might pay less attention in think or bonus trials (which merely require remembering the trial type and target location) than in track trials (which require actively tracking the target object in order to determine the target location). Subjects' equal levels of engagement during all trial types becomes critical when interpreting the neuroimaging results, as we want to ensure that

any differences we see in activation between attitude condition is not merely attributable to differences in task demands. Therefore, in addition to measuring overall response accuracy, we analyzed response accuracy just to the relevant question type for each trial. In the case of the track trials, the relevant question is, *"Is the [target object] in [some location]?"*. For the bonus trials, the relevant question is, *"Will you get a bonus if the [target object] is in [some location]?"* and for the think trials the relevant question is, *"Were you asked to think about the [target object] being in [some location]?"*

Evaluating accuracy for just the trial relevant questions allows us to compare attention across trial types. If subjects are paying less attention in think and bonus trials, they should produce fewer correct responses to the trial relevant questions for those trials than for track trials. Instead, we find that subjects perform very well on the bonus and think trial relevant questions (93.7% and 93.5% respectively) and somewhat more poorly on the track trial relevant questions (88.6% accuracy; see Figure 2). These results suggest that it is unlikely that subjects are paying more attention in track than bonus or think trials.

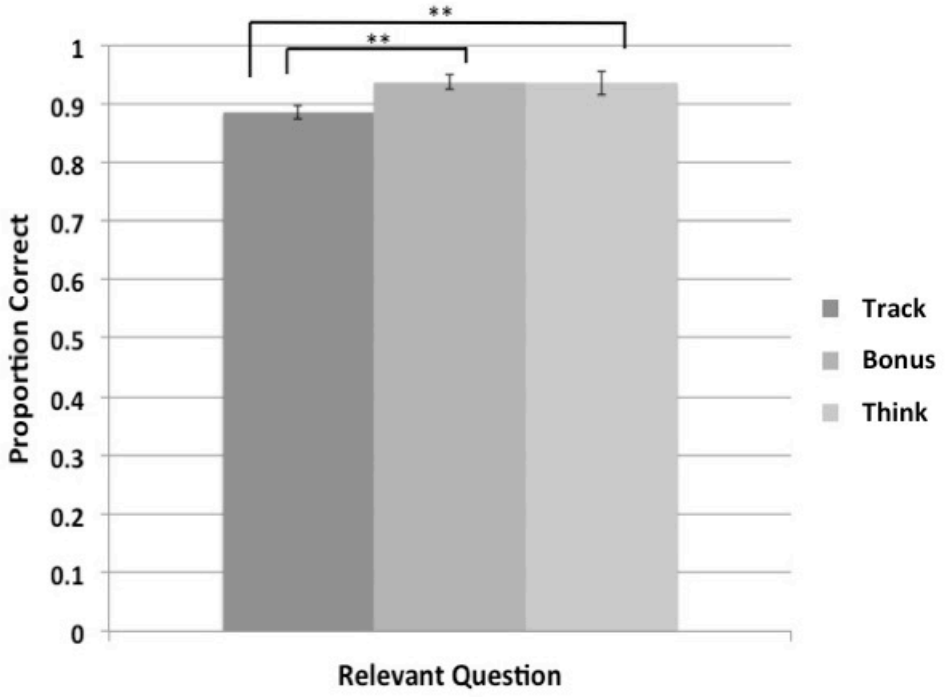


Figure 2. Response accuracies to task relevant questions. While there was no significant main effect of attitude, post-hoc analyses reveal a significant difference between the proportions of correct responses to task relevant questions in belief vs. desire trials and in belief vs. think trials. $^{**}p < .01$

III. PROPOSITIONAL ATTITUDES IN THE BRAIN

What happens in our brains when we believe that something is true, desire that something is true, or are merely thinking about something? To address this question we measured the neural activity associated with attitude condition (belief, desire, or think), across propositions, comparing levels of activation between attitudes.

General Methods

At the single subject level, blood-oxygenation-level-dependent (BOLD) responses during the critical delay period were estimated using a canonical hemodynamic response function with the event duration set to eight seconds (length of the delay period). An ordinary least squares regression was performed, with attitude condition as the primary regressors and motion parameters entered as regressors of no interest. General linear tests were done on the contrasts of interest (belief vs. desire, belief vs. think, desire vs. think). We then submitted the beta maps from the individual general linear tests to one-sample t-tests, producing group-level t-maps. We used Monte Carlo simulation to perform cluster-wise correction for multiple comparisons on the resulting t-maps with a voxel-wise significance threshold of $p < .001$ and a cluster-wide threshold of $p < .05$ (see Appendix for more information on image acquisition and preprocessing).

Desiring

Prior to identifying brain regions preferentially engaged when desiring a proposition, we sought to confirm that subjects desired the outcome in bonus trials more than in the other two trial types. One might imagine that there is some intrinsic reward associated with tracking the object correctly, leading subjects to desire the proposition in

belief trials perhaps as much as in desire trials, in which an extrinsic monetary bonus is given. To test this, we examined the neural responses to "successful" outcomes, either tracking the object correctly in a belief trial or having the target object end up in the target location yielding a $5 bonus in the desire trials. At the end of each trial, an arrow appears on the screen indicating the final location of the target object. On successful track trials, this arrow will point to what subjects believe to be the target object's final location. On winning bonus trials, this arrow will point to the pre-specified target location, indicating that the subject has won $5 on that trial. If subjects find trials in which they received a monetary bonus to be more rewarding than trials in which they were able to successfully track the target object, we would expect to see more activation in brain regions associated with reward during this arrow presentation in winning desire trials than in accurately tracked belief trials.

Using the univariate analysis procedures described above, we compared activation during the arrow presentation in winning bonus trials to activation during the same time period in accurately tracked belief trials. We found large swaths of increased brain activation for winning desire trials relative to accurately tracked belief trials bilaterally in the medial frontal lobe, anterior cingulate, insula, striatum, midbrain, thalamus, middle temporal gyrus, and inferior parietal lobule (see Figure 3). Reward processing consistently increases neural responses in the ventral striatum, medial prefrontal cortex, amydala, insula, and midbrain (Camara, Rodriguez-Fornells, Ye, & Münte, 2009; Düzel et al., 2009; McClure, York, & Montague, 2004). Many of the regions identified in our analysis are central to this established reward network. However, in order to ensure that these results were not merely the product of subjects paying less attention to the arrow

presentation in belief trials, we compared activation in winning bonus trials to losing bonus trials. We found regions of increased activation for winning bonus trials relative to losing trials that significantly overlapped with the contrast map for winning bonus trials over correctly tracked belief trials (see Figure 3). This suggests that these results are not merely a product of differentiable levels of attention but rather that relative to correctly tracking an object, subjects find winning a bonus to be much more rewarding.

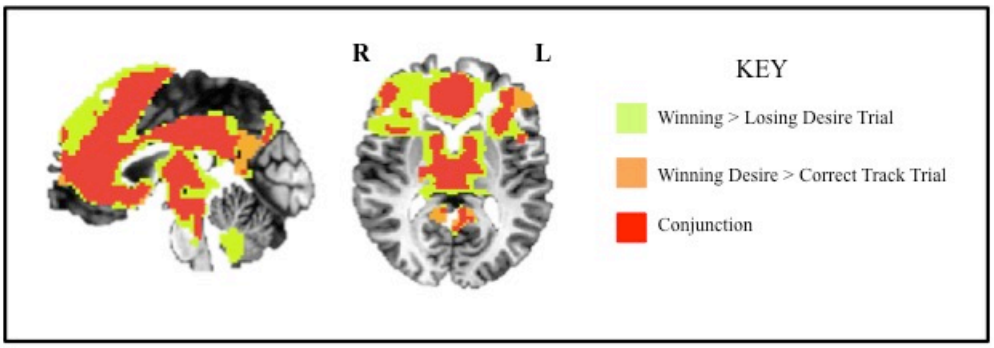


Figure 3. Neural activation during the arrow presentation at the end of each trial, indicating the final location of the target object. In desire trials, an arrow pointing to the target location indicates that the subject has won a $5 bonus. In belief trials, the arrow confirms whether or not the subject has correctly tracked the target object. Voxel-wise $p < .001$, cluster-wide corrected $p < .05$.

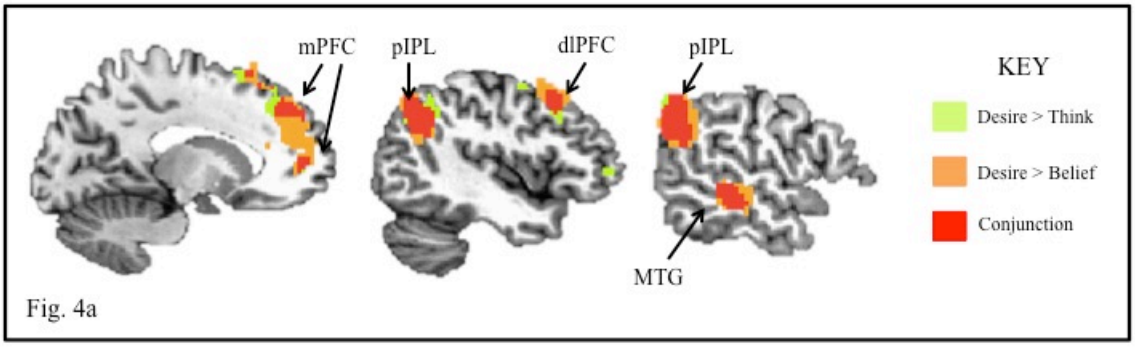
After establishing that subjects likely do find the winning bonus outcome to be more rewarding than other potentially rewarding trial outcomes, we sought to identify brain regions specifically associated with the desire for the propositional states that lead to these outcomes. To do this we searched the brain for regions that were significantly more active for desire trials relative to think trials and for desire trials relative to belief trials. We then performed a conjunction analysis to identify regions of overlap between the two contrasts. The conjunction of the desire over think and desire over belief contrasts revealed several areas of significant activation, including regions in the left medial and superior frontal gyri of the medial prefrontal cortex (mPFC), the left

dorsolateral prefrontal cortex (dlPFC), the left posterior inferior parietal lobule (pIPL), and the left middle temporal gyrus (MTG) (see Figure 4a and Appendix for full results).

Given that desire in our task is elicited by the anticipation of a rewarding outcome, we had expected to see neural activation associated with reward anticipation. However, we failed to find any increased activation in the classic reward or hedonic response regions such as the OFC, ventral striatum, or insula. Rather, many of the "desire" regions from our task, particularly the mPFC, the pIPL, and the MTG, fall within the default mode network. The default network is activated by internally oriented cognitive processes such as thinking about the past, imagining the future, or mind-wandering (Buckner, Andrews-Hanna, & Schacter, 2008). In general, this network works in opposition to the dorsal attention network, which is engaged during externally focused cognitive functions, especially those requiring attentional control (Fox, Corbetta, Snyder, Vincent, & Raichle, 2006). A key component of the dorsal attention network is the dlPFC, where we also found significantly more activation for desire than belief or think trials.

Although these two networks are thought to be anti-correlated (Fox, Zhang, Snyder, & Raichle, 2009), a pair of recent studies has found that both the default and dorsal attention networks are activated during mental simulations of future goal-directed action (Gerlach, Spreng, Gilmore, & Schacter, 2011; Gerlach, Spreng, Madore, & Schacter, 2014). In one such study, participants were asked to imagine both the steps they would take to reach a goal and events associated with having attained the goal. Compared to a control task, the goal-related simulations engaged a very similar collection of brain regions to those we identified here (see Figure 4b for a comparison), including the pIPL,

mPFC, MTG, and dlPFC (Gerlach et al., 2014). Interestingly, a more focused analysis in the Gerlach, 2014, study found that the dlPFC was functionally connected to the default network particularly when subjects were imagining the steps they would take to achieve a goal relative to imagining the goal outcome. In our study, subjects are unable to take any action to realize a specific desired outcome, however the dramatic overlap between our findings and Gerlach's, suggests that the impulse to engage in some kind of planning may persist, even in the absence of the ability to take action. This idea is consistent with a *world-to-mind* direction-of-fit view of desire in which desires function to drive us to act on the world to shape it to fit our mental states.



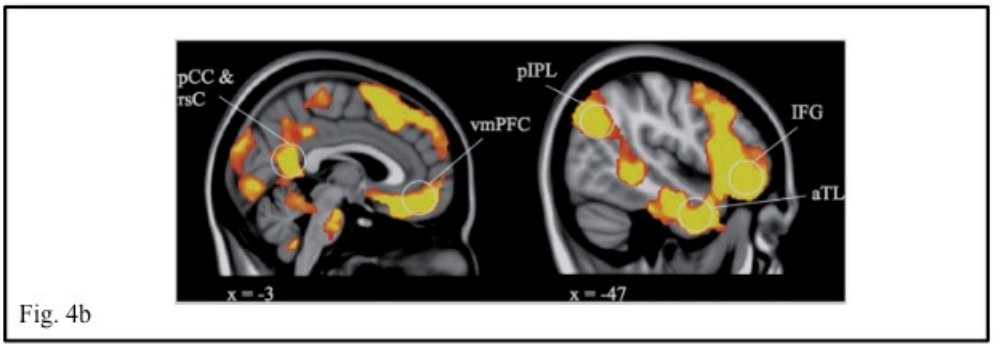


Figure 4. a) Greater neural activation associated with the desire condition relative to the belief and think conditions in the left mPFC, pIPL, dlPFC, and MTG. Voxel-wise $p < .001$, cluster-wide corrected $p < .05$. b) For comparison, activation in the default network for goal simulations from Gerlach et al., 2014).

Believing

In order to measure neural activation associated specifically with believing a proposition, we looked for brain regions exhibiting increased activation for belief trials relative to think trials and belief trials relative to desire trials. After correcting each contrast for multiple comparisons, we performed a conjunction analysis to identify regions of overlap. We found a single 160 voxel cluster in the right posterior parietal cortex (PPC), covering portions of the posterior superior parietal lobule (rSPL) and anterior precuneus, that was significantly more active in the belief condition than both the think and desire conditions (see Figure 5).

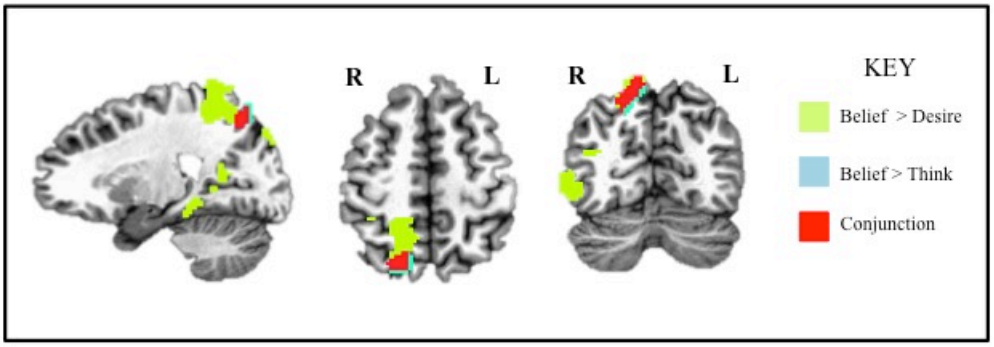


Figure 5. Greater neural activation associated with the belief condition relative to the desire and think conditions in the right PPC. Voxel-wise $p < .001$, cluster-wide corrected $p < .05$.

This region of the PPC appears to play an important role in mental imagery and visio-spatial working memory. In one study, subjects engaged in a two-back task that required keeping either objects or locations in working memory for short periods of time. They found increased activation in this region of the PPC when subjects were maintaining location information in working memory relative to when they were remembering objects (Schon, Tinaz, Somers, & Stern, 2008). A second study found that TMS to this particular region of the PPC impaired subjects' ability to mentally rotate letters relative to TMS to another region of the PPC (Pelgrims, Andres, & Olivier, 2009).

These findings suggest that this region of the PPC might be critical for the short-term maintenance and manipulation of visually represented information. We could imagine that in our task, when a proposition (that an object is in a specific location) is believed, subjects engage with the proposition more deeply, leading to increased mental imagery and engagement of their visuo-spatial working memory. This theory supports the Cartesian view of belief, suggesting that believing a proposition leads to extra processing, beyond merely representing the proposition.

Another collection of studies has implicated this region of the PPC in object tracking (Kimmig et al., 2008; Shulman et al., 1999). In particular, one such study found that activation in this region of interest (ROI) was specifically associated with the occulor-motor components of object tracking relative to viewing stationary stimuli (Kimmig et al., 2008). The role of our ROI in object tracking raises the possibility that the belief condition activation in the PPC is an artifact of the study design. In belief trials, subjects visually track the target object to its final location, while in the think and desire trials subjects are told the target location, and then observe the covered objects being shuffled, learning nothing about their final locations from the shuffle. Although all of our analyses are restricted to the eight-second delay period after the object tracking has been completed, it is possible that signal related to the shuffle/tracking period is still present during the delay period.

As described in the introduction, the objects were shuffled either by rotating the blue squares around the screen or by using arrows to point in the direction the blue squares would be moving. These two shuffle types were randomly assigned to each trial so that approximately 50% of the trials in each condition were of each shuffle type. If, in

fact, the neural activity from the tracking period is bleeding over into the delay period, causing more activation in the PPC for belief than think or desire trials, we might also expect to see activation associated with shuffle type (rotation versus arrows) in this same region during the delay period.

To test this, we searched the brain for regions whose mean levels of activation during the eight-second delay period differed between trials with the rotation versus arrow shuffles. In fact, we did find differential activation specifically in our PPC ROI for the rotation versus arrow shuffles. These results do suggest that we are finding some shuffle-associated activation during the eight-second delay period. However, if the increased activation in the PPC was purely a product of object tracking, we would expect to see more activation for the rotation rather than the arrow version of the object shuffle. Instead, activation increased in this ROI for trials with the arrow shuffle relative to those in which the covered objects were visibly rotated. One may interpret the arrows, more than the visible rotation, as inducing beliefs about unseen states of affairs, namely the new locations of the objects. Understood this way, the arrow-related activation from the shuffle period would, in fact, be consistent with the idea that this region plays a role in the encoding of beliefs about unseen states of affairs.

Thinking

Both the think and desire trials required entertaining a proposition without holding a clear belief about its truth. Therefore, in order to identify neural activation associated with thinking about a proposition, we looked for regions that showed increased activation for think trials, relative to belief trials, as well as increased activation

for desire trials, relative to belief trials. There were no regions that were significantly more active for think trials than belief trials after correcting for multiple comparisons. However, there was one 50 voxel cluster in the right inferior frontal gyrus (right IFG) that was significant at the uncorrected voxel-wise threshold of $p < .001$ and fell just below the required cluster size to reach cluster-wide significance. A large portion of this same region of the right IFG was significantly (corrected $p < .05$) more active for desire trials than belief trials (see Figure 6)[4] making this a strong candidate brain region to be associated specifically with thinking about a proposition without necessarily believing it.

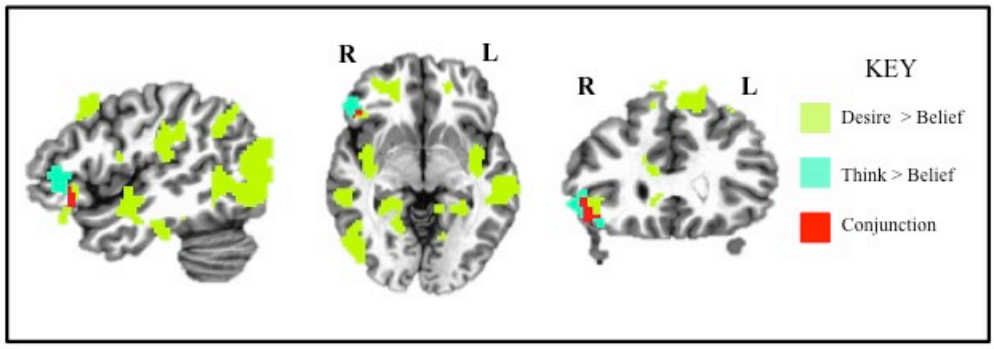


Figure 6. Greater neural activation associated with non-belief conditions relative to the belief condition in the right IFG. Desire > Belief contrast: voxel-wise $p < .005$, cluster-wide corrected $p < .05$. Think > Belief contrast: voxel-wise $p < .005$, 100 voxel cluster minimum.

Right IFG functioning is relatively heterogeneous. The region seems to play 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, 2007). The right IFG has also been identified as part of the mirror neuron system (Kilner, Neal, Weiskopf, Friston, & Frith, 2009) and as

[4] For the think>belief/desire>belief conjunction analysis, the t-maps from both contrasts were set at a voxel-wise threshold of $p < .005$. The desire>belief contrast was then corrected for multiple comparisons to a cluster-wide threshold $p < .05$. Because the think>belief contrast did not survive corrections for multiple comparisons, a minimum cluster size was set at 100 voxels.

such plays an important role in the experience of empathy (Shamay-Tsoory, Aharon-Peretz, & Perry, 2009).

However, the region has most consistently been implicated in response inhibition (see Aron et al., 2004, 2014 for reviews). The right IFG is a key region in the fronto-basal-ganglia inhibition network (Verbruggen & Logan, 2008), which also includes the presupplementary motor area (preSMA), the subthalamic nucleus (STN), and the globus pallidus. It is both functionally and structurally connected to the preSMA, and both the right IFG and preSMA are consistently activated during response inhibition across experimental tasks (Aron et al., 2007; Swann et al., 2012). Further, both regions are connected via white matter tracts to the STN, which also shows increased activation during inhibition (although it is not known whether the right IFG triggers STN activation directly or does so through the preSMA; Aron et al., 2007, 2014). When inhibitory processes are intended to stop a motor action, the STN, in turn, activates segments of the globus pallidus (Frank, 2006), which suppresses motor output.

The majority of the research on the right IFG and response inhibition has been done using the Go/No-go and stop-signal tasks (see Swick et al., 2008, for a meta-analysis). One study designed to tease apart the cognitive and neural sub-processes captured by these tasks, found response inhibition related activation specifically in our right IFG region of interest (Sebastian et al., 2013). This study utilized a Go/No-go-stop-signal task hybrid to distinguish between response inhibition required by stimulus incongruities (which the authors called “interference inhibition”), the ability to withhold a motor response (“action withholding”), or the ability to stop an already ongoing motor response (“action cancellation"). In the interference inhibition condition, subjects

received an arrow indicating which hand to use to push a response button on either the ipsilateral (congruent trials) or contralateral (incongruent trials) side of the screen. In the action withholding condition, subjects occasionally received a "no-go" cue to withhold their response on the same screen as button press cue. In the action cancellation condition, subjects received a "stop" cue to cancel their response after they had already received the button press cue. The authors found peak activation in our ROI for incongruent relative to congruent trials in the interference inhibition condition and for no-go trials relative to go trials in the action withholding condition. While they did find right IFG activation in the action cancelation condition, it did not overlap with the portion of the right IFG implicated in our study. These findings suggest that our right IFG ROI is more important for the cognitive rather than motor components of response inhibition.

While much of the research on the right IFG has been related to motor response inhibition specifically, it has also been found to play a significant role in other types of psychological inhibition. The right IFG is important for some aspects 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 or resolving stimulus conflict (Anderson & Greene, 2001; Egner, 2001; Mitchell et al., 2007). In one such study, researchers looked at the neural activation associated with conflict adaptation (Egner, 2011). Conflict adaptation occurs when subjects' responses improve to subsequent presentations of conflicting stimuli. In this study, they used a face-stroop task, where subjects had to identify the gender of a person

in a picture with either the word “male” or “female” written over it. In congruent trials, the word matched the gender in the picture and in incongruent trials it did not. The authors found increased activation in the right IFG, peaking in our “think” ROI, when subjects had an incongruent trial following a congruent trial relative to an incongruent trial following a congruent trial. In other words, activation in our right IFG ROI was highest when subjects did not have the opportunity to engage in conflict adaptation and were required to maximally engage cognitive control.

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. This is consistent with the Spinozan theory of belief, which posits that any thought that is entertained is automatically believed and then must be unbelieved through an extra processing step. As suggested by the aforementioned behavioral research, this extra processing step could be thought of as inhibiting some of the automatic belief processes (Gilbert et al., 1993). In fact one such study has found increased activation in the right IFG specifically when inhibiting beliefs in the context of deductive reasoning (Goel & Dolan, 2004). Although we cannot draw strong conclusions that the right IFG is playing an inhibitory role in our task, or that if it is it is inhibiting belief specifically, our results are consistent with both the neuroimaging research on inhibition and the behavioral research on belief automaticity.

Representationalists who endorse folk psychology posit that common sense references to beliefs and desires denote mental states that are implemented by consistent

and dissociable patterns of neural activity (Fodor, 1987; Schiffer, 1981). In support of this theory, we found clear activation associated with desire in portions of the default network (the left posterior inferior parietal lobule, medial temporal gyrus and medial prefrontal cortex) as well as in the left dorsolateral prefrontal cortex across a set of propositions. Likewise, we found activation associated with belief, across propositions, in the right posterior parietal cortex. This extra belief-related activation also supports the Cartesian view of belief, according to which believed propositions must be actively marked as believed, after being comprehended and assessed for veracity (Gilbert, 1991). The literature on this region of the PPC and evidence from our analysis of the activation associated with shuffle type, suggests that this ROI may be particularly engaged when imagining or visualizing believed, but unseen, states of the world. We also identified a region of increased activation in the right inferior frontal gyrus when propositions are not necessarily believed. These findings support the Spinozan view of belief, according to which, propositions are automatically believed when comprehended and extra processing – potentially inhibitory processing – is required to represent them as unbelieved (Gilbert, 1991).

IV. THE NEURAL REPRESENTATIONS OF BELIEVED AND DESIRED PROPOSITIONS

Having gathered some evidence for consistent and distinct neural activation associated with different propositional attitudes, we move now to understanding how the propositions themselves are represented in the brain. In the introduction, we outlined two different hypotheses about how proposition information is represented in the brain, both within and across propositional attitudes. In this section, we use multi-voxel pattern analysis to test these hypotheses by investigating how different propositional attitudes affect the neural representation of proposition information.

Testing the Stable Representation Hypothesis

The *stable representation hypothesis* posits that there are regions of the brain that contain consistent propositional content across propositional attitude. To test this, we used Multi-Voxel Pattern Analysis (MVPA). MVPA goes beyond evaluating differences in mean levels of brain activation to identify patterns of neural activity related to representational content (Haxby et al., 2001; T. M. Mitchell et al., 2008; Norman, Polyn, Detre, & Haxby, 2006). In this case, we used MVPA to identify brain regions that reliably carry information about the proposition independent of attitude. Due to power constraints, we began our analyses by looking for brain regions that contain information about either object or location across attitude. We then searched for regions that contain information about both the object and target location. If the *stable representation hypothesis* is correct, we would expect to see regions that contain information about which object and location are the targets of each trial, regardless of attitude condition.

Location

In order to identify brain regions in which we were able to decode the target location for each trial, we ran a whole brain searchlight analysis (Kriegeskorte and Bandettini (2007); see Appendix for details) using a naïve Bayes classifier to distinguish between each of the four locations (top-left, top-right, bottom-left, bottom-right). Because we are interested in location representation across attitudes, we trained our pattern classifier to identify which of the four locations was presented in each trial for two of the three attitude conditions and then tested the classifier's performance in trials for the third, cross-validating across each possible permutation of training and test attitudes. We then ran a t-test on the resulting accuracy maps, allowing us to identify brain regions, across subjects, where the classifier performed significantly better than chance. Our results were corrected for multiple comparisons at a voxel-wise threshold of $p < .005$ and a cluster-wide threshold of $p < .05$.

Our whole-brain search revealed significantly better than chance location decodability in a single 10,619-voxel cluster centered in the occipital lobe (see Appendix for detailed results). Medially, the cluster covers the entire lobe, extending from the occipital pole to the parieto-occipital fissure. Laterally the cluster reaches the bilateral posterior middle temporal gyrus, precuneus, and posterior fusiform (see Figure 7). The occipital lobe's functional specialization in visual processing (Clarke & Miklossy, 1990; DeYoe et al., 1996; Zeki et al., 1991) suggests that the representation of location in our task was almost certainly visual. Because the cluster encompasses the early visual cortex, which processes the basic components of visual stimuli, it is likely that participants were directing their gaze towards the target location. The extension of the cluster into the

precuneus indicates that participants may also have been engaging in some visual imagery (see Cavanna and Trimble (2006) for a review on the role of the precuneus in imagery). Given these results, it is likely that participants were looking at the target location, rather than processing the location in working memory, and as such may not have encoded location beyond the low-level visual representation.

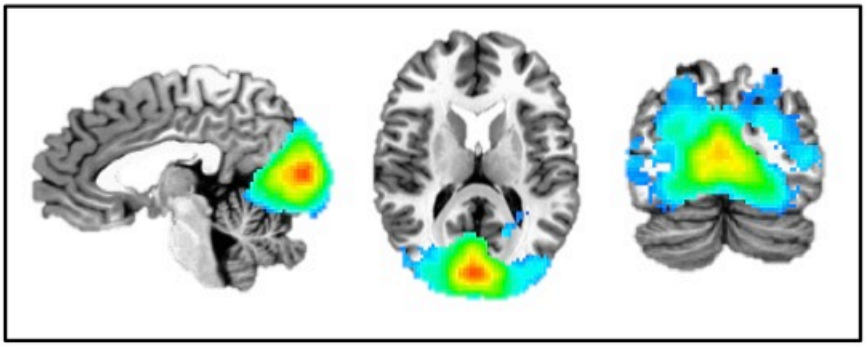


Figure 7. Location representation in the occipital lobe across all three attitude conditions. Voxel-wise $p <$ .005, cluster-wide corrected $p < .05$.

Object

Using the method for our location classification described in detail above, we performed a whole-brain searchlight analysis using a pattern classifier to identify the target object (mop, dog, snake, hose) in each trial, across attitudes. Our object classification analysis revealed large clusters in the left parahippocampal gyrus, bilateral superior temporal gyri, bilateral frontal gyri, and bilateral insula where our object classification performed significantly above chance. There were also significant clusters of object decodability in the left superior parietal lobule, right pre- and post-central gyrus, and bilateral precuneus (see Appendix for detailed results).

One shortcoming of this particular analysis is that in order for the classifier to succeed, the object needs to be decodable in only two out of three attitude conditions. For example, if the classifier is able to successfully identify object in belief and desire trials

only, and one of these is the test condition, the classifier may be able to succeed by sufficiently learning the patterns of activation associated with each object from the trials in the belief condition only, without exploiting object information from the think condition. To address this issue, we performed a follow-up analysis. First, we used the t-map generated by the whole-brain object classification as a mask, limiting our further investigation to just these regions.[5] Then, within these regions specifically, we ran three separate classifications, one with trials from each attitude condition as the test set. In other words, we ran one classification where we trained the classifier to identify the target object in belief and desire trials and then tested it on think trials, a separate classification where we trained only on think and desire trials and tested on belief trials, and a third classification where we trained the classifier on belief and think trials and then tested on desire trials. We then performed t-tests on each set of classification maps to find group level significance. Finally, we applied a 4.4mm smoothing to each t-map (equivalent to 2 voxels) and identified regions of overlap, where object classification was significantly better than chance (at a voxel-wise and corrected cluster-wide threshold of $p < .05$) for all three attitude conditions.

Our conjunction analysis found reliable object decodability in a 168 voxel cluster in the left parahippocampal gyrus (PHG) for the overlap of each of the three attitude conditions (see Figure 8). Very small clusters were also found in the left cerebellum (35 voxels) and left fusiform gyrus (20 voxels). In the PHG, the classifier's mean accuracy

[5] We masked our analyses with the significant clusters from our whole-brain object classification because we are interested in what factors were driving the significant effect in these regions in particular. However, just as with other post-hoc tests, corrections for multiple comparisons should be applied to these analyses. Because we have not applied corrections to the results of these analyses, they should be interpreted as preliminary.

for each attitude type was within .5% indicating that the classifier is performing similarly well, regardless of which attitude was used as the test condition.

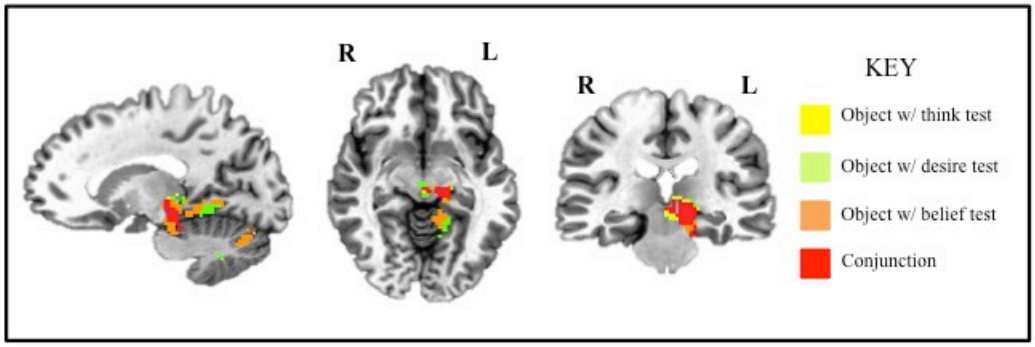


Figure 8. Object representation in the right PHG across all three attitude conditions. Voxel-wise $p < .05$, cluster-wide corrected $p < .05$.

The PHG is most often engaged by tasks that involve integrating pieces of spatial information in an environment (Aminoff, Kveraga, & Bar, 2013). Most relevant to the present study is the significant role the PHG plays in encoding and retrieving object-location associations (Committeri et al., 2004; Janzen & Van Turennout, 2004; Maguire et al., 1998; Malkova & Mishkin, 2003; Milner, Johnsrude, & Crane, 1997; Owen, Milner, Petrides, & Evans, 1996; Sommer, Rose, Weiller, & Büchel, 2005). However, in the present study, we were able to reliably decode object, but not location information in the PHG. As described earlier, the results of our location classification were confined primarily to the visual cortex. This purely visual neural activation likely results from the subjects looking towards the target location, and may be the source of our inability to find decodable location information in the PHG.

While it is most often implicated in studies of associative memory involving spatial processing, the PHG also seems important for other kinds of object-related associative processing, including the association between two objects across time (Hales, Israel, Swann, & Brewer, 2009) and across category (Düzel et al., 2003). In fact, one

recent study suggests that our specific PHG region of interest is involved in the retrieval of non-spatial episodic memories (Hoscheidt, Nadel, Payne, & Ryan, 2010). In this study subjects were asked to recall either spatial episodic memories (for example the spatial relationships between people or items at a recent event) or non-spatial episodic memories (details of a recent event not related to locations or spatial relationships). Relative to control, they found more activation for non-spatial rather than spatial memory retrieval in our PHG ROI.

Object-Location Conjunction

While identifying regions that contain object or location information is an important component of capturing task-relevant proposition representation, the attitudes in this study are not about objects or locations alone, but rather a specific object *being in* a specific location. Subjects don't merely desire *dog*, but rather *dog in the upper right corner*. Therefore, we sought to identify brain regions that contain information about the object-location conjunctions. In this case, merely running a sixteen-way classification distinguishing between each object-location combination is insufficient. A classifier that is able to identify just object or just location may be able to perform better than chance at this task. Instead we ran separate classifications for object in each location and location for each object, and then looked for brain regions in which both types of classifications performed significantly better than chance.

To do this we split the entire dataset first by object, resulting in 36 each of dog, mop, snake, and hose trials. Then, within each of those object-specific trial sets, we used a pattern classifier to identify the target location for each trial. Using a leave-one-out

procedure, the classifier was trained to identify location for two of the three attitudes and then tested on the third repeating this procedure for every iteration of train and test conditions. Using a whole-brain searchlight analysis, we produced a location classification accuracy for every voxel in the brain for each object-specific trial set (therefore, we had four location classification accuracies for each voxel). We then averaged the accuracies in each voxel together, creating an average location classification accuracy across objects for each voxel. Finally, we performed t-tests to identify clusters of voxels whose average classification accuracies were significantly above chance at the group level. We repeated the same procedures for our object classification, running four separate classifications (one for each location), averaging the object classification accuracies across locations, and identifying regions where the group-wide average object classification accuracy was significantly better than chance. Finally we looked for regions of overlap between our location (averaged across objects) and object (averaged across locations) significant t-maps.[6]

Using this procedure, we were unable to identify any regions that contained information about both object (averaged across locations) and location (averaged across objects). There are several possibilities why this might be the case. First, this analysis was likely very underpowered. Splitting the dataset up by object (or location) reduces the total number of trials used in each analysis from 144 to 36, leaving only 24 training and 12 test trials for each round of classification. Classification accuracy decreases as sample size, and especially the number of training samples, decreases (Raudys & Jain, 1991).

[6] Prior to performing the conjunction analysis, t-maps were corrected for multiple comparisons at a voxel-wise threshold of $p < .005$ and cluster-wide threshold of $p < .05$.

Second, it is possible, as suggested by the results of the location classification, that during the delay period subjects are turning their gaze towards the target location instead of actively maintaining the location information in working memory. In this case, they might be fully encoding the target object, but not the target location, decreasing the likelihood that we would find overlap in the neural representation of object and location. Finally, it may be that there are no brain regions that contain the conjunctively bound object-location information. In this case, object and location may be represented in disparate brain regions that are bound through temporal synchrony, where object and location are integrated through the synchronization of discrete regions that separately encode object or location information (Singer & Gray, 1995).

While we were unable to find a representation of the object-location conjunction, our analysis revealed clear instantiations of both object and location information across attitude. This finding provides clear support for the *stable representation hypothesis*, that at least some aspects of proposition representation are consistent, regardless of whether the proposition is believed, desired, or merely being thought about.

Testing the Variable Representation Hypothesis

The *variable representation hypothesis* posits that propositional attitude affects the representation of the proposition. If the variable representation hypothesis is correct, we might expect to find brain regions that contain more decodable propositional content in one attitude condition than another. To test this we used MVPA to isolate brain regions that carry information about the proposition within each attitude condition and then

compared object decodability in one attitude condition relative to another. Because of our concerns about the depth of encoding of location information, we limited this analysis to object representation only.

We began by dividing the data by attitude condition, yielding 48 trials for each attitude condition. We then ran separate object pattern classifications within each attitude condition. Using a whole-brain searchlight analysis and a leave-one-out cross-validation procedure, we trained our classifier to identify which of the four objects was presented in each trial for 11 of 12 runs, then tested the classifier's performance in the 12^{th} run, repeating this process for every permutation of training and test runs. These analyses provided us with whole-brain maps of the object classifier's accuracy for each attitude condition for each subject. To assess differences in object representation between attitude conditions, for each subject we subtracted the voxel-wise accuracies for one attitude condition from the voxel-wise accuracies for a second attitude condition (for example object classification accuracy in belief trials minus object classification accuracy in think trials). Finally we used t-tests to find clusters of voxels with significant between-attitude difference scores at the group level. These results were corrected for multiple comparisons at voxel-wise $p < .005$ and cluster-wide $p < .05$ levels.

The effect of desire on object representation

To identify differences in object representation related to desiring, we contrasted object classification accuracies between the desire and think conditions and between the desire and belief conditions. We find several regions exhibiting significantly better object decodability for desire trials than for think trials in the left putamen, cuneus/precuneus,

cingulate, and medial frontal gyrus/paracentral lobule (see Figure 9 and Appendix for full results). In these regions, object classification accuracy is better than chance in desire trials, and at or below chance in think trials. Further, in each of these regions we see chance levels of object decodability in belief trials, although the difference in classification accuracy between desire and belief trials is not significant.[7]

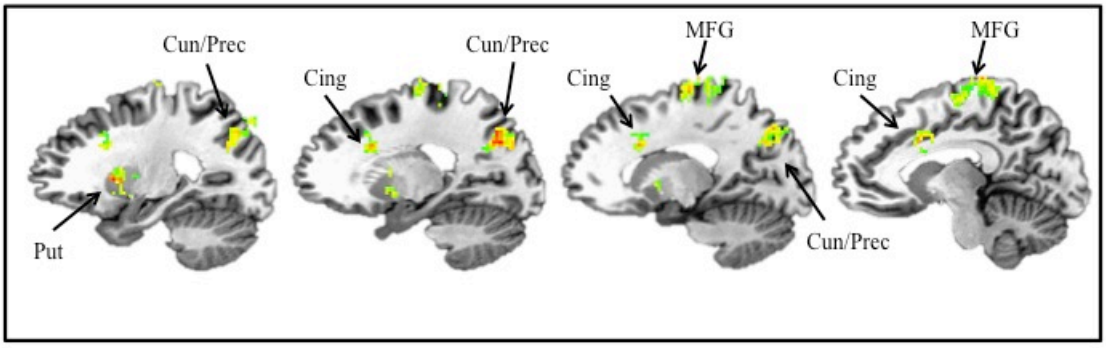


Figure 9. Regions supporting significantly better object decoding in desire trials than in think trials. Put = putamen, Cun = cuneus, Prec = precuneus, Cing = cingulate, MFG = medial frontal gyrus. Voxel-wise $p < .005$, cluster-wide corrected $p < .05$.

Given its established relationship to reward anticipation, the increased object decodability in the left putamen is of special interest. In fact, several studies have found associations between the ROI identified here and reward anticipation (Kirsch et al., 2003; Knutson, Adams, Fong, & Hommer, 2001; Knutson, Fong, Adams, Varner, & Hommer, 2001). However, with our task we did not find increased mean levels of activation for the desire condition. Rather, here we find information about *what* is being desired, when it is being desired. The body of research most closely related to our findings is on the neural

[7] While the difference in object classification accuracy between desire and think conditions was significant at a corrected $p < .05$, some of this effect was driven by the fact that the classifier performed slightly below chance in each of these regions in think trials. Object classification accuracy in desire trials (when *not* being compared to think trials) in each of these regions is significant at a voxel-wise threshold of $p < .005$, but does not survive cluster-wide corrections for multiple comparisons.

representation of rule- or task-reward associations. Generally, these studies ask participants to follow a rule or complete a task in order to receive a future reward. For example, in Reverbi et al.'s (2012) paper, participants were required to push left and right buttons in a specific order depending on the stimuli being presented. In rule- and task-reward association studies like this, researchers then use multi-voxel pattern analysis to identify brain regions that reliably carry information about the particular rule or task associated with a reward. Most often, task or rule information is identifiable in the lateral prefrontal cortex (Bengtsson, Haynes, Sakai, Buckley, & Passingham, 2009; Bode & Haynes, 2009; Bunge, Kahn, Wallis, Miller, & Wagner, 2003; Reverberi et al., 2012; Sakai & Passingham, 2003; Woolgar, Hampshire, Thompson, & Duncan, 2011) although one study in monkeys has also found that neurons in the striatum respond differentially to rule information.

Critically, in our study, we are not cuing a task that must be performed to receive a reward, but rather a condition or state of the world (proposition), which if true, will yield a reward. In the studies described above, subjects do not associate the reward with any particular task or rule itself, but rather with their successful completion of the task or adherence to the rule. In our study, reward is associated explicitly with a specific state of the world holding true, ultimately making the proposition itself the target of desire. One very recent study has found that the identity of a rewarding cue is decodable in the striatum (Anderson, 2016). In this study, subjects were asked to indicate whether a line in either a red or green circle is horizontal or vertical. Correct responses were given a monetary reward. The authors found that during the reward period of each trial (after subjects had made their horizontal or vertical selection) they were able to reliably decode

the color of the preceding cue in the caudate tail. This study may be more similar to the present work. Anderson finds that the caudate carries the identity of the rewarding stimuli much in the same way we find information about the potentially rewarding stimuli in the putamen. However, they are decoding the cue identity during the receipt of the reward, not while subjects are actively anticipating, or hoping for the reward.

While its role in the striatal reward system makes a compelling case for why we might see object information in the putamen when the object is a constituent of a desired proposition, there is reason to believe the precuneus and middle frontal gyrus might also be important for the representation of desired objects. The precuneus and middle frontal gyrus both fall in the default network (Buckner et al., 2008; Raichle et al., 2001). As discussed in the previous chapter, we also found increased activation in portions of the default network, specifically the mPFC, the pIPL, and the MTG, for the desire condition, relative to the think and belief conditions. In conjunction, these findings suggest that distinct regions of the default network may be associated with engaging in sustained desire, and with representing the content of those desires.

The effect of belief on object representation

To evaluate whether object representation is modulated by propositional attitude, we contrasted object classification accuracies in the belief condition with classification accuracies in the think and desire conditions. We found significantly better object decodability in the belief condition relative to the think condition in the right posterior superior temporal gyrus (pSTG), right paracentral lobule, and bilateral cuneus (see Figure 10 and Appendix for full statistics). In these regions, object classification accuracy is

better than chance in belief trials, and at or below chance in think trials. Further, in each of these regions we see chance levels of object decodability in desire trials, although the difference in classification accuracy between belief and desire trials is not significant.[8]

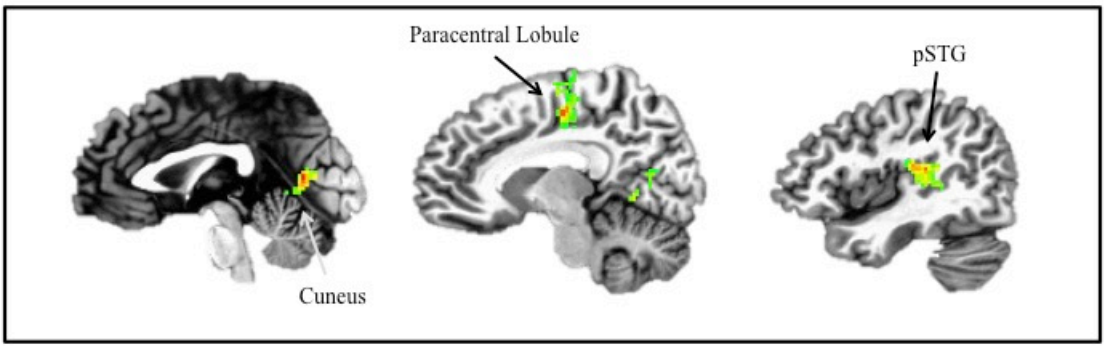


Figure 10. Regions supporting significantly better object decoding in belief trials than in think trials. Voxel-wise $p < .005$, cluster-wide corrected $p < .05$.

Given its place on the primary motor and somatosensory cortices (specifically in the regions serving the leg and foot; Fix, 2002), increased object decodability for belief trials in the right paracentral lobule is difficult to interpret. As part of the primary visual cortex, increased object decodability in the cuneus during belief trials may indicate that individuals are engaging in more mental imagery in the belief condition. However, several studies have shown that the increased activation in the cuneus is associated only with the perception of stimuli but not engaging in mental imagery (Ganis, Thompson, & Kosslyn, 2004; Roland & Gulyás, 1995).

The right pSTG is functionally heterogeneous. Our pSTG region of interest at

[8] As with the prior analysis, some of the difference between object classification accuracy in belief versus think trials was driven by the fact that accuracy in think trials tended to be slightly below chance. Object classification accuracy in belief trials (when *not* being compared to think trials) in each of these regions is significant at a voxel-wise threshold of $p < .005$, but does not survive cluster-wide corrections for multiple comparisons.

least partially overlaps with the auditory cortex, and as such, activation in the region is often associated with auditory processing, frequently playing a role in speech perception (Evans et al., 2014; Zaehle, Geiser, Alter, Jancke, & Meyer, 2008). However, this region has also been implicated in word retrieval and object identification (Garn, Allen, & Larsen, 2009; Murtha, Chertkow, Beauregard, & Evans, 1999). In one PET study, subjects were asked to identify an animal based on a line drawing, or make a semantic judgment about that animal (does it have hooves, claws, or neither). They found more activation in our right pSTG region of interest for both conditions relative to baseline (Murtha et al., 1999). However, because object identification is inherent to making semantic judgments about that object in the aforementioned study, it is unclear whether the increased activation in the semantic judgment condition was due to the judgment itself or the process of identifying the object in order to make the judgment. In a more recent fMRI study (Pihlajamäki et al., 2005), subjects were presented with five objects placed on a grid. In the novel location condition, subjects were shown the same five objects, one of which was moved to a new location. In the novel object condition, subjects were shown five objects in the original locations, but one object was replaced with a novel object. The authors found increased activation in our pSTG ROI when subjects were viewing a novel object, relative to when they were viewing a novel location. Taken together, this body research is consistent with our interpretation of the pSTG as especially important for the representation of object identity when the object is a constituent of a believed proposition.

In our prior analysis on the neural activation associated with propositional attitude, we found evidence for both the Cartesian and Spinozan views of belief.

Consistent with the Cartesian view we found increased activity for belief trials in the right PPC (though possibly related to prior object tracking). Under this view, it is possible that the increased object decodability in belief trials in right pSTG (and perhaps the cuneus and paracentral lobule) reflects some further encoding of believed propositions. However, consistent with the Spinozan view, our univariate analyses also revealed increased activity in the right IFG, which is associated with inhibition, when propositions are comprehended but not believed. When viewed through a Spinozan lens, the better object decodability we find in the right pSTG in belief trials may not reflect the engagement of an additional belief process. Rather, it may be a consequence of a dampening or distribution of the proposition representation when propositions are *not* believed.

With respect to the interaction between propositional attitudes and the representation of propositional content, we proposed two theories. According to the *stable representation hypothesis*, the neural representation of a proposition is invariant across propositional attitudes. According to the *variable representation hypothesis*, the representation of the proposition changes depending on the operative propositional attitude. Here, too, we find evidence supporting both hypotheses. Across attitudes, we found consistent location information in the occipital lobe and consistent object information in the parahippocampal gyrus. These findings are consistent with the *stable representation hypothesis*. However, consistent with the *variable representation hypothesis*, we identified regions in which decoding success varied with propositional attitude. Object decoding was significantly better in the right pSTG, paracentral lobule,

and bilateral cuneus when the object was a constituent of a proposition that was being believed. Likewise, object decoding was significantly better in the left putamen, cuneus/precuneus, cingulate, and medial frontal gyrus/paracentral lobule when the object was a constituent of a proposition that was desired.

One interpretation of the Language of Thought Hypothesis (Fodor, 1987) posits that the representations of believed propositions are instantiated in a figurative "belief box," while the representation of desired propositions are instantiated in a figurative "desire box" (Schiffer, 1981). In its most literal form, this theory suggests that we would find one or more brain regions that represent propositional information when and only when those propositions are believed, and a distinct set of regions that represent propositional information when and only when those propositions are desired. Our evidence in support of the *variable representation hypothesis* is, for now, partly consistent with this view. In the regions in which we find significantly better object decodability for desire than think trials, we do not merely find that object representation is moderated by attitude, with some level of object decodability in belief and think trials and a higher degree of decodability in desire trials. Rather, we find a complete absence of decodable object information in think and belief trials, with the classifier performing at or below chance. Likewise, in the regions in which we are significantly better able to decode object information in belief than think trials, we are unable to successfully classify object in think and desire trials.

The absence of decodable object information in these regions for some propositional attitudes may simply be due to our methodological limitations. Our most successful decoding of object information is only 5.51% above chance in belief trials and

5.48% above chance in desire trials. These relatively near chance levels of object decodability make it likely that any less successful decoding will be at chance. And, as noted above, there is positive evidence for the representation of propositional content that is stable across attitudes in at least one other region. Nevertheless, the data are, for now, consistent with the intriguing idea that the brain, in addition to having stable representations of propositional content, has something like a "belief box" and a "desire box," i.e. regions that represent propositional content when and only when that content is part of a believed/desired proposition.

V. CONNECTING PROPOSITIONAL ATTITUDES TO PROPOSITIONS

So far, we have identified several brain regions preferentially engaged when desiring a proposition, a single region that is preferentially engaged when thinking about a proposition, and (somewhat tentatively) a region that is preferentially engaged when believing that a proposition is true. What's more, we have identified neural representations of propositions that are invariant across propositional attitudes as well as representations that are modulated by propositional attitude, and possibly unique to a specific propositional attitude. But what is the relationship between the attitude-associated brain regions and the propositional content bearing regions? In other words, how does the brain connect propositions to the attitudes we hold about them?

While there is still a tremendous amount of work to be done on this front, we have taken an exploratory first step towards investigating this relationship using functional connectivity. Functional connectivity is the temporal association of neural activation in disparate brain regions (Friston, 2011; Van Den Heuvel & Pol, 2010) and is most commonly performed by assessing correlations in the time-series data for two or more brain regions (Biswal, Zerrin Yetkin, Haughton, & Hyde, 1995; Friston et al., 1994). Here we use functional connectivity to look at the relationship between the brain regions associated with a particular attitude, and those associated with the attitude dependent proposition representation.

Functional Connectivity Methods

ROI Generation

In the analyses discussed in the previous chapters we found dissociable sets of brain regions that are preferentially more active depending on whether propositions were

believed, desired, or merely being thought about. Further, we found regions in which we are better able to decode propositional content (specifically object information) depending on the attitude condition. In the set of analyses discussed here, we explored the relationships between these two sets of regions. More specifically, we examined the correlations between the mean level of activation in the attitude-specific regions, and the probability of correctly classifying object in the regions of attitude dependent object decodability.

Because we do not have a full independent set of data for establishing the ROIs for these analyses, we used a leave-one-out method to established unbiased ROIs for each subject, generated by the group effects from the other subjects. Using 28 of the 29 subjects eligible for this analysis (one subject was excluded because of alignment issues), ROIs were generated using the univariate and multivariate analysis procedures described in chapters three and four above. All voxels meeting a $p < .05$ uncorrected threshold and falling within a minimum 50 voxel cluster were included as potential ROIs. This process was repeated with each subject left out, creating 29 sets of N-1 subject ROIs.

Functional Connectivity Analysis

For each subject we extracted their pre-processed data, with the global signal removed, in just the ROIs using the masks created with the other 28 subjects. The data for each voxel in each ROI were averaged across the four TRs comprising the eight-second delay period (adjusted for hemodynamic lag), leaving one data-point per trial, per voxel. In the attitude specific ROIs the data were then averaged again across voxel, providing a per-trial average activation for each attitude-specific ROI. For the object representation

ROIs, we followed the data analysis procedures described in chapter four, training a classifier to identify the target object for each trial across propositional attitudes. We extracted the posterior probability of choosing the correct object during each classification test phase, then averaged the posterior probabilities for each test attitude. This analysis gave us a single average posterior probability of choosing the correct object for each trial in each ROI. Finally, for each subject, we correlated the trial-by-trial mean activation in select attitude-specific ROIs with the average trial-by-trial posterior probability in select object representation ROIs, separately for each attitude condition. We then performed one-sample t-tests on each set of correlation coefficients to find significant group-wide correlations. Given the exploratory nature of these analyses, none of the results were corrected for multiple comparisons.

Thinking About Propositions

Our univariate analyses found increased activation in the right inferior frontal gyrus when subjects were in either the think or desire conditions relative to the belief condition. Given the widespread implication of the right IFG in inhibitory processes (see Aron et al., 2004; 2014 for reviews), we hypothesized that this region plays a role in inhibiting automatic belief. This idea is consistent with the Spinozan view of belief, according to which any proposition that is entertained is automatically believed and then, if its veracity is in doubt, must be actively "unbelieved" (Gilbert, 1991). If the right IFG is, in fact, suppressing beliefs about propositions, we might expect to see a negative association between the activation in the right IFG and the representation of propositional

content in regions that more reliably encode proposition information when a proposition is being believed.

To test this hypothesis, we examined the correlation between the mean level of activation in the right IFG and the posterior probability of correctly classifying object in the three regions in which we were better able to decode object information in belief trials as opposed to think trials. These regions are: the right posterior superior temporal gyrus (pSTG), the right paracentral lobule, and the bilateral cuneus. In fact, we found a significant negative association between activation in the right IFG and object decodability in the right pSTG for think trials ($t(28) = -2.37$, $p = .025$). Across subjects, the average correlation between the mean level of activation in the right IFG and the posterior probability of correctly classifying the target object in the pSTG, specifically in think trials, was -.06. While these small within-subject correlations were very small, the coefficients across subjects tended to be negative, with a mean value consistently less than zero. Thus, while weak, the negative correlation between activation in the right IFG and object decodability in the right pSTG, specifically in think trials, appears to be reliable.

Given that this result is not corrected for multiple comparisons, it should be treated as very preliminary. However, this finding does lend credence to the idea that the pSTG, among the set of regions identified as those containing more decodable object information in belief than think trials, might be especially important for distinguishing believed propositions from those that are not believed. Although they are merely correlational, these results also support the theory that activation in the right IFG has a dampening effect on the proposition representation when the proposition is not believed.

However, more work must be done before we are able to conclude that activation in the IFG is, in fact, *causing* the decrease in object decodability in the pSTG.

Desiring Propositions

When comparing object decodability in desire and think conditions, we found that the left putamen, left cuneus/precuneus, left cingulate, and left medial frontal gyrus/paracentral lobule contained more object information in desire conditions than think conditions. We suggested that because of its role in the reward network, the putamen is a strong candidate for representing object information when those objects are associated with future rewards. We also argued that given that we have found more desire-related activation in some parts of the default network, we might expect to also see regions of the default network, such as the precuneus or medial frontal gyrus, contain information about *what* is being desired. In order to test whether any of these regions do in fact interact with those associated with desiring a proposition, we correlated the average activation in our regions that were selectively activated for desire trials (the left medial prefrontal cortex, dorsolateral prefrontal cortex, middle temporal gyrus, and posterior inferior parietal lobule) with the posterior probability of correctly classifying object in the left putamen, precuneus, and medial frontal gyrus. We found that in desire trials in particular, there was a weak, but consistent, correlation across subjects between activation in the left medial prefrontal cortex and the classifier's posterior probability in the left medial frontal gyrus. Although the within subject correlations are once again very small, the mean correlation coefficient (average r = .07) across all 29 subjects was significantly above zero ($t(28)=2.47$, p =.02). Additionally we found significant

relationships between the posterior probability correctly classifying object in the left precuneus and mean activation in the left dorsolateral prefrontal cortex (r=.05, t(28)=2.23, p = .034) and in the left posterior inferior parietal lobule (r=.05, t(28)=2.12, p = .043) for think trials only.

Once again, because these results are not corrected for multiple comparisons, they are far from definitive. However, the potential for a relationship between activation in the medial prefrontal cortex and object classification in the middle frontal gyrus lends support to the theory the default network may play an important role in engaging in sustained desire for a specific outcome. Given its importance in prospection (Buckner & Carroll, 2007; Spreng & Grady, 2010), it is possible that when engaging in sustained desire, subjects are imaging a future in which the object of their desires comes to fruition.

Using functional connectivity, we have begun to explore the relationship between brain regions associated with propositional attitude, and those that preferentially represent object information when the object is a constituent of a proposition that is being believed or desired. Our preliminary results suggest that when a proposition is not necessarily believed, activation in the right IFG is negatively associated with object decodability in the right pSTG. One possible explanation for these results is that the IFG is playing a role in unbelieving automatically believed propositions and, subsequently, dampening the proposition representation in the pSTG. We also find that when a proposition is being desired, activation in the left MPFC is associated with object decodability in the left MFG. This is consistent with the theory that the default network plays a central role in desiring a specific outcome. Given that these findings are both

uncorrected and correlative, we cannot make strong claims about the nature of the relationships presented here. However, we hope to use these findings as a starting point in the development of a causal model for *how* the neural instantiation of attitude affects the representation of propositions.

VI. GENERAL DISCUSSION AND CONCLUSIONS

General Discussion

The ultimate goal of this project is to understand how propositions, ideas about possible realities, and propositional attitudes, such as believing and desiring, are flexibly combined in the brain. To this end, we sought to identify neural activity associated with believing, desiring, and merely thinking about a limited set of propositions, and whether propositional attitudes affect the neural representation of these propositions. To do this we designed a novel shell game task in which subjects were induced to believe, desire, or merely think about one of four objects being in one of four locations. By having subjects complete this task while undergoing functional neuroimaging, we were able to study the neural activity associated with distinct propositional attitudes when applied to the same propositions (object-location combinations), propositional content both within and across propositional attitudes, and the relationships between neural activity related to propositional attitudes and patterns of activity encoding propositional content.

Consistent with a folk psychological view of propositional attitudes (Fodor, 1987), we hypothesized that we would find consistent and dissociable regions of increased neural activation for specific propositional attitudes, independent of propositional content. Commensurate with this hypothesis, when comparing activation in the desire condition to activation in the belief and think conditions, we found increased activation in the left MPFC, pIPL, MTG, and dlPFC. Interestingly, most of these regions (MPFC, pIPL, and MTG) fall within the default network, which shows increased activation during prospection (Buckner et al., 2008; Spreng & Grady, 2010). Further, research has shown that activation in the default network and dlPFC occurs in conjunction specifically when individuals are planning and imagining the steps they will

take to achieve a goal (Gerlach et al., 2014). These findings suggest that in the desire condition, the default network is playing a role in attaching values to relatively complex imagined states of affairs.

We outlined two theories concerning the neural representation of belief (Gilbert, 1991). The Cartesian view posits that propositions are first represented, then assessed for veracity. Those that are deemed true are then believed. This view predicts some kind of increased neural activation corresponding to the believing of the proposition beyond the mere representation of the proposition. Alternatively, the Spinozan view posits that all propositions are automatically believed, and then if deemed to be untrue or of unknown veracity, an additional "unbelieving" process occurs. This theory predicts additional neural activation associated with thinking about a proposition while not believing it, as compared to belief. Prior behavioral and neuroimaging evidence suggests that this process of unbelieving requires cognitive control (Gilbert et al., 1993; Harris et al., 2009; Harris et al., 2008; Hawkins & Hoch, 1992; Marques et al., 2009).

When comparing activation associated with our belief condition relative to the non-belief conditions (think and desire), we found increased activation in the right PPC. Consistent with the Cartesian view of belief, this increased PPC activation may be indicative of some additional processing that occurs when a proposition is believed. Further, prior research on this region of the PPC and evidence from our analyses of the activation associated with shuffle type, suggests that this ROI may be particularly engaged when imagining or visualizing believed, but unseen, states of the world.

To test the Spinozan theory, we made the reverse contrast, looking for increased activity in non-belief (desire and think) trials relative to belief trials. Here, we found

increased activation in the right IFG. This increased activation for non-believed relative to believed propositions, is consistent with the idea that propositions are automatically believed merely through comprehension and then, if deemed to be not true (or if the truth value is unknown), "unbelieved." Further, given the central role of the right IFG in inhibition (see Aron et al., 2004; 2014 for reviews), we suggest that this process of unbelieving may be an inhibitory process, where cognitive control is required to actively suppress the prepotent belief response.

In addition to examining the neural bases of distinct propositional attitudes, we asked whether propositional attitudes influence the representation of the propositions themselves. On the one hand, it is possible that the neural representation of a proposition is consistent across attitude. If this *stable representation hypothesis* is true, we would expect to find brain regions that contain reliably decodable proposition information across attitude. However, it is also possible that a proposition's representation changes depending on whether the proposition is being believed or desired – what we call the *variable representation hypothesis*. To test these hypotheses, we used multi-voxel pattern analyses to identify brain regions encoding object or location information, both across and within attitude conditions.

Consistent with the *stable representation hypothesis*, we were able to identify the target location, across attitudes, at above chance levels in the visual cortex. Further, we were able to identify the target object above chance, across attitudes, in the right parahippocampal gyrus. However, we were unable to reliably decode object-location conjunctions above chance in any brain region. Given that the location representation was constrained to the visual cortex, it is likely that subjects were looking at or attending to

the target location in each trial, using this visual shortcut to allow them to bypass deeper encoding of the target location. This might partially explain why we were able to decode target object, but not the target location, in the right parahippocampal gyrus, a region commonly implicated in integrating object/location information (Committeri et al., 2004; Janzen & Van Turennout, 2004; Maguire et al., 1998; Malkova & Mishkin, 2003; Milner et al., 1997; Owen, Milner, et al., 1996; Sommer et al., 2005). Likewise, if subjects did not use working memory to encode the target location, this may explain why we were unable to decode the conjunctively bound object-location in any other region. However, it is also possible that object-location binding occurs not through a conjunctive representation, but rather another process such as through temporal synchrony (Singer & Gray, 1995).

To test the *variable representation hypothesis* we looked for brain regions where we were better able to decode the target object in one attitude condition relative to another. We found significantly better object decodability for desire trials in the left putamen, cuneus/precuneus, cingulate, and MFG/paracentral lobule for desire trials when compared to think trials. Although we did not find increased mean levels of activation in the putamen for desire trials as we might expect given its role in the reward network (Kirsch et al., 2003; Knutson, Adams, et al., 2001; Knutson, Fong, et al., 2001), to the extent of our knowledge, this is the first study in which information about the identity of a reward-related object is decoded in the putamen.

We observed increased activity within several default network regions for the desire condition, relative to the think and belief conditions. As constituents of the default network, finding decodable object information (when the object is part of a desired

proposition) in the precuneus and MFG fits with the supposition that this network is important for imaging desired future outcomes. Further bolstering this theory, in an exploratory analysis we found a significant (uncorrected) correlation between activation in the left MPFC and object decodability in the left MFG, suggesting that separate regions of the network may work in concert to connect the motivational processes that constitute desiring to the objects of those desires.

When comparing belief to think trials we found significantly better object decodability in the right pSTG, right paracentral lobule, and bilateral cuneus. The right pSTG in particular has been implicated in representing object identification, especially when the object name or identity must be maintained in working memory (Garn et al., 2009; Murtha et al., 1999; Pihlajamäki et al., 2005). We suggested that this belief-specific object decodability may be indicative of additional processing that occurs when a proposition is believed. When a proposition is not believed, on the other hand, we might find that this extra processing must be inhibited, resulting in a dampening of the proposition representation in this region. Consistent with this theory, in think trials only, we find a significant (uncorrected) negative correlation between activation in the right IFG (which is preferentially engaged when propositions are not believed) and object decodability in the right pSTG. In other words, when a proposition is not believed, more activation in the right IFG is associated with poorer object decodability—and presumably reduced encoding—in the right pSTG. These findings also support our hypothesis that the right IFG is inhibiting proposition belief, potentially by moderating the representation of the proposition in the right pSTG.

Limitations

The primary objective of the shell game task was to induce the first-order experience of belief, desire, and thinking about over a range of propositions. However, we found that this was not possible without some potential confounds. In order to create conditions where subjects either had a clear belief or no belief about the final object-location combination, we generated easily trackable (for the belief condition) and untrackable (for the desire and think conditions) object shuffles. In the belief trials, subjects deduced the final object location by tracking the target object to that location. On the other hand, because the tracking was impossible in the desire and think conditions, subjects were presented with the target location in writing immediately before the shuffling occurred.

Knowing the potential for the trackability of the shuffle to influence our results, we took several steps to mitigate the impact of potential confounds. First, we created two different methods of shuffling, either by having the squares physically move around the screen or by having arrows indicate the proceeding location of the squares. Both of these shuffling methods occurred in all three attitude conditions and we collapsed across shuffling method in our analyses. Second, we performed all of our analyses during the eight-second delay period following the object tracking. During this time, subjects received the same visual input despite attitude condition or the tracking method. However, we were unable to control for the fact that in the belief condition only, subjects tracked the target object to the final location, while in the desire and think conditions they were given a written target location. Ultimately, it is possible that the difference in either the style or role of object tracking between conditions had an effect on our results

indicating increased activation in the right PPC in the belief condition. As discussed previously in this paper, our belief related activation was in a region associated with object tracking and we did find differential activation in this region for our arrows versus rotation shuffle methods.

During our analyses we found decodable location information only in the visual cortex. This suggests that subjects were looking at the target location, allowing them to forgo any further processing of the location information. Unfortunately, this may have prevented us from identifying representations of object-location conjunctions. It also suggests that we may not have captured location representation at a level that that meets the standard of being able to be flexibly combined with different propositional attitudes. To address this, we excluded location from our within-attitude and connectivity analyses. Future versions of this experiment should ensure that subjects fully encode the target location and integrate that information with both the target object and the propositional attitudes. One way to address this may be to present both the target object and target location information only in writing, requiring that subjects generate some linguistic representation of the proposition information. Alternatively, future experiments using this paradigm might include instructions to subjects to fix their gaze on a central point and/or eye tracking to monitor gaze direction.

Our results indicate that unlike the target location representation, the target object representation in our study was not merely sensory, as we observed representations of object identity in regions that are far removed from primary sensory cortices. Nor does it seem to reflect subjects' merely remembering the target object without integrating the object with their beliefs or desires about the propositional outcome (the target objects'

being in a specific location). There are several pieces of evidence that suggest that this might be the case. First, the right parahippocampal gyrus, where we found stable object representation, is often associated with conceptual object representation, as well as with integrating the representation of object with other pieces of conceptual information (Düzel et al., 2003; Hales et al., 2009). Additionally, consistent with our *variable representation hypothesis,* we found that object representation is modified by propositional attitude. Further, this attitude specific object representation appears to be functionally connected to the specific belief and desire related activation. Nevertheless, in future analyses we could take additional steps towards addressing this concern. By further investigating which features of the target object our pattern classifiers are exploiting in order to succeed at object classification we can discern more information about the level of object encoding. Our object stimuli (the dog, mop, snake, and hose) were designed so that we have two pairs of visually matched stimuli (dog/mop and snake/hose) and two pairs of conceptually matched stimuli (dog/snake and mop/hose). If we find that, in regions such as the right PHG, the target object is represented as a member of its category (animal vs. tool) rather than based on its visual features (green and long vs. white and stringy), this suggests that object is being represented at a level beyond the merely low-level sensory representation.

In this study, we used functional connectivity to examine the relationship between processes associated with specific propositional attitudes and the encoding of propositional content. Although our functional connectivity analyses did produce several hypothesis driven significant results, they do not survive correction for multiple

comparisons and are therefore very preliminary. Because of the limitations of our current dataset, this issue may only be able to be addressed with a replication.

Future Directions

We performed our functional connectivity analyses with the ultimate goal of developing a model for *how* the brain flexibly combines propositions with propositional attitudes such as believing and desiring. But what might this model look like? An obvious next step is to use similar exploratory functional connectivity analyses to begin to unpack the relationship between attitude-invariant object decodability in the right PHG, attitude specific object decodability in the right pSTG and left MFG, and attitude associated activation in the left MPFC and right IFG. The ultimate goal, then, would be to use the results of these analyses to develop a model that fully integrates the neural activation associated with each attitude with both the attitude invariant and attitude specific proposition representations.

In our study, we find regions in which the representations of propositional content are stable across attitudes and others in which proposition representations appear to depend on attitude. However, we evaluated only a very limited set of object-location propositions, and the proposition information is primarily presented visually. These two factors likely constrain the representation of the proposition and potentially of the propositional attitude. Future studies should evaluate the effect of attitude on the representation of these propositions when the information is presented either in writing or auditorily. Additionally, we would like to expand our set of propositions beyond these concrete object-location combinations. Humans are able to entertain an effectively

infinite number of propositions, many of which are quite abstract. But what is the relationship between the representations of the propositions, *there is a god*, or, *dividing by zero is undefined*, and the attitudes we hold about them? Is a belief or desire about an abstract proposition, like *there is a god,* the same kind of belief or desire we hold about concrete propositions like *the dog is in upper right hand corner*? In our future studies we hope to incorporate both concrete and abstract propositions to begin to answer some of these questions.

If propositional attitudes are defined by their *direction-of-fit* (Anscombe, 1957; Searle, 1983), it follows that desires, with a *world-to-mind* fit, are meant to drive us to shape the world to match our mental states. As such, desires are affective motivators that produce behaviors that tend toward the attainment of a desired outcome. However, in our present study, participants are given no opportunity to engage in desire driven behaviors to shape the outcome. Rather, they can merely *hope* the target object lands in the desired location. Is the neural representation of hope, without opportunity for action, the same as desire that drives behavior towards achieving some goal? How does the neural activation associated with hope or desire, and with the proposition that is desired, engage with neural processes associated with goal oriented action planning? In future studies we would like to investigate the neural mechanisms by which desires and beliefs combine to produce goal-directed behavior.

Conclusions

Philosophers such as Fodor (1987) and Dretske (1988) argue that our folk psychological view of the mind is essentially accurate – that the things we call "beliefs",

"desires", "hopes", "fears", etc. are not just convenient fictions, but real things, robust mental entities that neuroscience will ultimately describe and explain, rather than displace. This Representational Theory of Mind claims that we have in our minds, and ultimately our brains, mental relations (propositional attitudes) and mental representations of possible states of the world (propositions) that flexibly combine and recombine in a compositional fashion to produce both logical sequences of private thought and observable, intentional behavior (Fodor, 1981, 1987; Schiffer, 1981; Sterelny, 1990). Up until this point, there has been no systematic attempt to characterize these mental processes and representations in neural terms and to understand how the neural instantiation of propositional attitudes relates to the neural representation of propositional content. Consistent with the Representational Theory of Mind, we find evidence for distinct neural processes associated with believing, desiring, and merely thinking about propositions. We also find patterns of activity that represent propositional content independent of whether the proposition is believed, desired, or merely thought about. Finally, by identifying regions in which propositional content is differentially represented depending on propositional attitude, and by associating proposition representation in these regions with regions preferentially engaged by different attitudes, we have begun to understand the neural mechanisms that enable us to believe, desire, or merely think about different ways the world could be.

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*(2), 790-795.

Allport, G. W. (1935). Attitudes. In C. Murchinson (Ed.), *A Handbook of Social Psychology* (pp. 798-844). Worcester, MA: Clark University Press.

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in cognitive sciences, 17*(8), 379-390.

Anderson, A. K., Christoff, K., Stappen, I., Panitz, D., Ghahremani, D., Glover, G., . . . Sobel, N. (2003). Dissociated neural representations of intensity and valence in human olfaction. *Nature neuroscience, 6*(2), 196-202.

Anderson, A. K., & Sobel, N. (2003). Dissociating intensity from valence as sensory inputs to emotion. *Neuron, 39*(4), 581-583.

Anderson, B. (2016). Reward processing in the value-driven attention network: reward signals tracking cue identity and location. *Social cognitive and affective neuroscience*, nsw141.

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410*(6826), 366-369.

Anscombe, G. E. M. (1957). *Intention*. Ithaca New York: Cornell University Press.

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*(14), 3743-3752.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in cognitive sciences, 8*(4), 170-177.

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*(4), 177-185.

Audi, R. (1994). Dispositional beliefs and dispositions to believe. *Noûs, 28*(4), 419-434.

Aydede, M. (2015). The Language of Thought Hypothesis. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015 ed.): Metaphysics Research Lab, Stanford University.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences, 7*(2), 84-91.

Begg, I. M., 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*(4), 446.

Bengtsson, S. L., Haynes, J.-D., Sakai, K., Buckley, M. J., & Passingham, R. E. (2009). The representation of abstract task rules in the human prefrontal cortex. *Cerebral Cortex, 19*(8), 1929-1936.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience, 21*(8), 2793-2798.

Berridge, K. C. (2009). 'Liking'and 'wanting'food rewards: brain substrates and roles in eating disorders. *Physiology & behavior, 97*(5), 537-550.
Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward:'liking','wanting', and learning. *Current opinion in pharmacology, 9*(1), 65-73.
Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine, 34*(4), 537-541.
Blood, A. J., Zatorre, R. J., Bermudez, P., & Evans, A. C. (1999). Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nature neuroscience, 2*(4), 382-387.
Bode, S., & Haynes, J.-D. (2009). Decoding sequential stages of task preparation in the human brain. *Neuroimage, 45*(2), 606-613.
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), 1-38.
Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in cognitive sciences, 11*(2), 49-57.
Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K., & Wagner, A. D. (2003). Neural circuits subserving the retrieval and maintenance of abstract rules. *Journal of neurophysiology, 90*(5), 3419-3428.
Camara, E., Rodriguez-Fornells, A., Ye, Z., & Münte, T. F. (2009). Reward networks in the brain as captured by connectivity measures. *Frontiers in neuroscience, 3*, 34.
Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296*(5576), 2191-2191.
Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain, 129*(3), 564-583.
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*(3), 607-623.
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*(39), 12315-12320.
Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *the Journal of Philosophy, 78*(2), 67-90.
Clarke, S., & Miklossy, J. (1990). Occipital cortex in man: Organization of callosal connections, related myelo-and cytoarchitecture, and putative boundaries of functional visual areas. *Journal of Comparative Neurology, 298*(2), 188-214.
Cole, M. W., & Schneider, W. (2007). The cognitive control network: integrated cortical regions with dissociable functions. *Neuroimage, 37*(1), 343-360.
Collins, J. (1988). Belief, desire, and revision. *Mind, 97*(387), 333-342.
Committeri, G., Galati, G., Paradis, A.-L., Pizzamiglio, L., Berthoz, A., & LeBihan, D. (2004). Reference frames for spatial cognition: different brain areas are

involved in viewer-, object-, and landmark-centered judgments about object location. *Journal of cognitive neuroscience, 16*(9), 1517-1535.
Cornman, J. W. (1968). On the Elimination of'Sensations' and Sensations. *The Review of Metaphysics*, 15-35.
Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage, 19*(2), 261-270.
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*(2), 537-545.
Critchley, H. D., & Rolls, E. T. (1996). Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *Journal of neurophysiology, 75*(4), 1673-1686.
Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in cognitive sciences, 7*(9), 415-423.
Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience, 23*(8), 3423-3431.
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*(2), 238-257.
Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of neurophysiology, 84*(6), 3072-3077.
Dennett, D. (1978). *Brainstorms*. Cambridge, MA: Bradford Books.
DeYoe, E. A., Carman, G. J., Bandettini, P., Glickman, S., Wieser, J., Cox, R., . . . Neitz, J. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences, 93*(6), 2382-2386.
Dolcos, F., & McCarthy, G. (2006). Brain systems mediating cognitive interference by emotional distraction. *Journal of Neuroscience, 26*(7), 2072-2079.
Dreher, J.-C., Kohn, P., & Berman, K. F. (2006). Neural coding of distinct statistical properties of reward information in humans. *Cerebral Cortex, 16*(4), 561-573.
Dretske, F. (1988). The explanatory role of content.
Düzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B. H., & Tobler, P. N. (2009). Functional imaging of the human dopaminergic midbrain. *Trends in neurosciences, 32*(6), 321-328.
Düzel, E., Habib, R., Rotte, M., Guderian, S., Tulving, E., & Heinze, H.-J. (2003). Human hippocampal and parahippocampal activity during visual associative recognition memory for spatial and nonspatial stimulus configurations. *Journal of Neuroscience, 23*(28), 9439-9444.
Egan, F. (1991). Propositional Attitudes and the Language of Thought. *Canadian Journal of Philosophy, 21*(3), 379-388.
Egner, T. (2011). Right ventrolateral prefrontal cortex mediates individual differences in conflict-driven cognitive control. *Journal of cognitive neuroscience, 23*(12), 3903-3913.

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature, 425*(6954), 184-188.
Elliott, R., Newman, J. L., Longe, O. A., & Deakin, J. W. (2003). Differential response patterns in the striatum and orbitofrontal cortex to financial reward in humans: a parametric functional magnetic resonance imaging study. *Journal of Neuroscience, 23*(1), 303-307.
Evans, S., Kyong, J., Rosen, S., Golestani, N., Warren, J., McGettigan, C., . . . Scott, S. (2014). The pathways for intelligible speech: multivariate and univariate perspectives. *Cerebral Cortex, 24*(9), 2350-2361.
Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33*(25), 10552-10558.
Feyerabend, P. (1963). Materialism and the mind-body problem. *The Review of Metaphysics*, 49-66.
Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*(5614), 1898-1902.
Fix, J. D. (2002). *Neuroanatomy*: Lippincott Williams & Wilkins.
Fodor, J. A. (1975). *The language of thought* (Vol. 5): Harvard University Press.
Fodor, J. A. (1981). Representations: Philosophical essays on the foundations of cognitive science.
Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*: The MIT Press.
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, 103*(26), 10046-10051.
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*(6), 3270-3283.
Frank, M. J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks, 19*(8), 1120-1136.
Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity, 1*(1), 13-36.
Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping, 2*(4), 189-210.
Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research, 20*(2), 226-241.
Garn, C. L., Allen, M. D., & Larsen, J. D. (2009). An fMRI study of sex differences in brain activation during object naming. *Cortex, 45*(5), 610-618.
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*(4), 1816-1824.

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*, nsu001.
Gilbert, D. T. (1991). How mental systems believe. *American psychologist, 46*(2), 107.
Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of personality and social psychology, 65*(2), 221.
Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortexin inductive and deductive reasoning. *Cognition, 93*(3), B109-B121.
Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2002). Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *Journal of Neuroscience, 22*(24), 10829-10837.
Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision research, 41*(10), 1409-1422.
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, 98*(7), 4259-4264.
Hales, J. B., Israel, S. L., Swann, N. C., & Brewer, J. B. (2009). Dissociation of frontal and medial temporal lobe activity in maintenance and binding of sequentially presented paired associates. *Journal of cognitive neuroscience, 21*(7), 1244-1254.
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*(22), 5623-5630.
Harris, S., Kaplan, J. T., Curiel, A., Bookheimer, S. Y., Iacoboni, M., & Cohen, M. S. (2009). The neural correlates of religious and nonreligious belief. *PLoS One, 4*(10), e7272.
Harris, S., Sheth, S. A., & Cohen, M. S. (2008). Functional neuroimaging of belief, disbelief, and uncertainty. *Annals of neurology, 63*(2), 141-147.
Hart, J., & Gordon, B. (1990). Delineation of single-word semantic comprehension deficits in aphasia, with anatomical correlation. *Annals of neurology, 27*(3), 226-231.
Hartwigsen, G., Price, C. J., Baumgaertner, A., Geiss, G., Koehnke, M., Ulmer, S., & Siebner, H. R. (2010). The right posterior inferior frontal gyrus contributes to phonological word decisions in the healthy brain: evidence from dual-site TMS. *Neuropsychologia, 48*(10), 3155-3163.
Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of consumer research, 19*(2), 212-225.
Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.
Holland, P. C., & Gallagher, M. (2004). Amygdala–frontal interactions and reward expectancy. *Current opinion in neurobiology, 14*(2), 148-155.

Hoscheidt, S. M., Nadel, L., Payne, J., & Ryan, L. (2010). Hippocampal activation during retrieval of spatial context from episodic and semantic memory. *Behavioural brain research, 212*(2), 121-132.
Humberstone, I. L. (1992). Direction of fit. *Mind, 101*(401), 59-83.
Janzen, G., & Van Turennout, M. (2004). Selective neural representation of objects relevant for navigation. *Nature neuroscience, 7*(6), 673-677.
Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature neuroscience, 10*(12), 1625-1633.
Kiefer, M. (2005). Repetition-priming modulates category-related effects on event-related potentials: further evidence for multiple cortical semantic systems. *Journal of cognitive neuroscience, 17*(2), 199-211.
Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., & Hoenig, K. (2008). The sound of concepts: four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience, 28*(47), 12224-12230.
Kiefer, M., Sim, E.-J., Liebich, S., Hauk, O., & Tanaka, J. (2007). Experience-dependent plasticity of conceptual representations in human sensory-motor areas. *Journal of cognitive neuroscience, 19*(3), 525-542.
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*(32), 10153-10159.
Kim, S. H., & Hamann, S. (2007). Neural correlates of positive and negative emotion regulation. *Journal of cognitive neuroscience, 19*(5), 776-798.
Kimmig, H., Ohlendorf, S., Speck, O., Sprenger, A., Rutschmann, R., Haller, S., & Greenlee, M. (2008). fMRI evidence for sensorimotor transformations in human cortex during smooth pursuit eye movements. *Neuropsychologia, 46*(8), 2203-2213.
Kirsch, P., Schienle, A., Stark, R., Sammer, G., Blecker, C., Walter, B., . . . Vaitl, D. (2003). Anticipation of reward in a nonaversive differential conditioning paradigm and the brain reward system:: an event-related fMRI study. *Neuroimage, 20*(2), 1086-1095.
Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci, 21*(16), RC159.
Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., & Hommer, D. (2001). Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport, 12*(17), 3683-3687.
Knutson, B., & Greer, S. M. (2008). Anticipatory affect: neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363*(1511), 3771-3786.
Kolb, B., Whishaw, I. Q., & Teskey, G. C. (2014). *An introduction to brain and behavior* (Vol. 1273).
Koob, G. F. (1992). *Dopamine, addiction and reward.* Paper presented at the Seminars in Neuroscience.
Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage, 38*(4), 649-662.

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural representation of subjective value under risk and ambiguity. *Journal of neurophysiology, 103*(2), 1036-1047.
Lewis, D. (1988). Desire as belief. *Mind, 97*(387), 323-332.
Lin, A., Adolphs, R., & Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Social cognitive and affective neuroscience, 7*(3), 274-281.
Lycan, W. G., & Pappas, G. S. (1972). What is eliminative materialism? *Australasian Journal of Philosophy, 50*(2), 149-159.
Maguire, E. A., Burgess, N., Donnett, J. G., Frackowiak, R. S., Frith, C. D., & O'keefe, J. (1998). Knowing where and getting there: a human navigation network. *Science, 280*(5365), 921-924.
Malkova, L., & Mishkin, M. (2003). One-trial memory for object-place associations after separate lesions of hippocampus and posterior parahippocampal region in the monkey. *Journal of Neuroscience, 23*(5), 1956-1965.
Marques, J. F., Canessa, N., & Cappa, S. (2009). Neural differences in the processing of true and false sentences: Insights into the nature of 'truth'in language comprehension. *Cortex, 45*(6), 759-768.
Mazoyer, B., Zago, L., Mellet, E., Bricogne, S., Etard, O., Houdé, O., . . . Tzourio-Mazoyer, N. (2001). Cortical networks for working memory and executive functions sustain the conscious resting state in man. *Brain research bulletin, 54*(3), 287-298.
McClure, S. M., York, M. K., & Montague, P. R. (2004). The neural substrates of reward processing in humans: the modern role of FMRI. *The Neuroscientist, 10*(3), 260-268.
McGrath, M. (2014). Propositions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 ed.): Metaphysics Research Lab, Stanford University.
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.
Milner, B., Johnsrude, I., & Crane, J. (1997). Right medial temporal–lobe contribution to object–location memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 352*(1360), 1469-1474.
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*(4), 292-297.
Mitchell, K. J., Johnson, M. K., Raye, C. L., & D'Esposito, M. (2000). fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory. *Cognitive Brain Research, 10*(1), 197-206.
Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191-1195.
Morris, J. S., Frith, C. D., Perrett, D. I., & Rowland, D. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383*(6603), 812.

Mummery, C. J., Patterson, K., Wise, R. J., Vandenbergh, R., Price, C., & Hodges, J. (1999). Disrupted temporal lobe connections in semantic dementia. *Brain, 122*(1), 61-73.

Murtha, S., Chertkow, H., Beauregard, M., & Evans, A. (1999). The neural substrate of picture naming. *Journal of cognitive neuroscience, 11*(4), 399-423.

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*(2), 241-268.

Nishijo, H., Ono, T., & Nishino, H. (1988). Single neuron responses in amygdala of alert monkey during complex sensory stimulation with affective significance. *Journal of Neuroscience, 8*(10), 3570-3583.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences, 10*(9), 424-430.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *Neuroimage, 11*(5), 424-446.

O'Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., & McGlone, F. (2001). Representation of pleasant and aversive taste in the human brain. *Journal of neurophysiology, 85*(3), 1315-1321.

O'Neill, M., & Schultz, W. (2010). Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron, 68*(4), 789-800.

O'Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., McGlone, F., Kobal, G., . . . Ahne, G. (2000). Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. *Neuroreport, 11*(4), 893-897.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current opinion in neurobiology, 14*(6), 769-776.

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., & Gross, J. J. (2004). For better or for worse: neural systems supporting the cognitive down-and up-regulation of negative emotion. *Neuroimage, 23*(2), 483-499.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology, 47*(6), 419.

Olson, I. R., Page, K., Moore, K. S., Chatterjee, A., & Verfaellie, M. (2006). Working memory for conjunctions relies on the medial temporal lobe. *Journal of Neuroscience, 26*(17), 4596-4601.

Owen, A. M., Doyon, J., Petrides, M., & Evans, A. C. (1996). Planning and spatial working memory: a positron emission tomography study in humans. *European Journal of Neuroscience, 8*(2), 353-364.

Owen, A. M., Milner, B., Petrides, M., & Evans, A. C. (1996). A specific role for the right parahippocampal gyrus in the retrieval of object-location: A positron emission tomography study. *Journal of cognitive neuroscience, 8*(6), 588-602.

Pelgrims, B., Andres, M., & Olivier, E. (2009). Double dissociation between motor and visual imagery in the posterior parietal cortex. *Cerebral Cortex*, bhn248.
Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature, 442*(7106), 1042-1045.
Pfaus, J., Damsma, G., Nomikos, G. G., Wenkstern, D., Blaha, C., Phillips, A., & Fibiger, H. (1990). Sexual behavior enhances central dopamine transmission in the male rat. *Brain research, 530*(2), 345-348.
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*(3), 210-219.
Piekema, C., Kessels, R. P., Rijpkema, M., & Fernández, G. (2009). The hippocampus supports encoding of between-domain associations within working memory. *Learning & Memory, 16*(4), 231-234.
Pihlajamäki, M., Tanila, H., Könönen, M., Hänninen, T., Aronen, H. J., & Soininen, H. (2005). Distinct and overlapping fMRI activation networks for processing of novel identities and locations of objects. *European Journal of Neuroscience, 22*(8), 2095-2105.
Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience, 27*(37), 9984-9988.
Postle, B., Stern, C., Rosen, B., & Corkin, S. (2000). An fMRI investigation of cortical contributions to spatial and nonspatial visual working memory. *Neuroimage, 11*(5), 409-423.
Price, H. (1989). Defending desire-as-belief. *Mind, 98*(389), 119-127.
Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological bulletin, 80*(1), 1.
Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(2), 676-682.
Ramsey, W., Stich, S., & Garan, J. (1990). Connectionism, eliminativism, and the future of folk psychology. *Philosophical Perspectives, 4*, 499-533.
Ranganath, C., & D'Esposito, M. (2005). Directing the mind's eye: prefrontal, inferior and medial temporal mechanisms for visual working memory. *Current opinion in neurobiology, 15*(2), 175-182.
Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence, 13*(3), 252-264.
Reverberi, C., Görgen, K., & Haynes, J.-D. (2012). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex, 22*(6), 1237-1246.
Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex, 15*(8), 1261-1269.

Roland, P., & Gulyás, B. (1995). Visual memory, visual imagery, and visual recognition of large field patterns by the human brain: functional anatomy by positron emission tomography. *Cerebral Cortex, 5*(1), 79-93.
Rolls, E. T., & Xiang, J.-Z. (2005). Reward-spatial view representations and learning in the primate hippocampus. *Journal of Neuroscience, 25*(26), 6167-6174.
Rorty, R. (1965). Mind-body identity, privacy, and categories. *The Review of Metaphysics*, 24-54.
Rota, G., Sitaram, R., Veit, R., Erb, M., Weiskopf, N., Dogil, G., & Birbaumer, N. (2009). Self-regulation of regional cortical activity using real-time fMRI: The right inferior frontal gyrus and linguistic processing. *Human brain mapping, 30*(5), 1605-1614.
Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1988). *Parallel distributed processing* (Vol. 1): IEEE.
Sakai, K., & Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature neuroscience, 6*(1), 75-81.
Schiffer, S. (1981). Truth and the theory of content. In H. Parrett & J. Bouverese (Eds.), *Meaning and Understanding* (pp. 205-224). Berlin: Walter de Gruyter.
Schon, K., Tinaz, S., Somers, D. C., & Stern, C. E. (2008). Delayed match to object or place: an event-related fMRI study of short-term stimulus maintenance and the role of stimulus pre-exposure. *Neuroimage, 39*(2), 857-872.
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599.
Schwitzgebel, E. (2015). Belief. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 ed.): Metaphysics Research Lab, Stanford University.
Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge, United Kingdom: Cambridge University Press.
Sebastian, A., Pohl, M., Klöppel, S., Feige, B., Lange, T., Stahl, C., . . . Tüscher, O. (2013). Disentangling common and specific neural subprocesses of response inhibition. *Neuroimage, 64*, 601-615.
Sellars, W. (1956). Empiricism and the Philosophy of Mind. *Minnesota studies in the philosophy of science, 1*(19), 253-329.
Sellars, W. (1963). Science, perception, and reality.
Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience, 27*(18), 4826-4831.
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*(3), 617-627.
Sharp, D. J., Scott, S., K., & Wise, R. J. S. (2004). Retreiving meaning after temporal lobe infraction: The role of the basal language area. *Annals of neurology, 56*, 836-846.
Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., & Petersen, S. E. (1997). Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. *Journal of cognitive neuroscience, 9*(5), 648-663.

Shulman, G. L., Ollinger, J. M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Petersen, S. E., & Corbetta, M. (1999). Areas involved in encoding and applying directional expectations to moving objects. *Journal of Neuroscience, 19*(21), 9480-9496.

Simanova, I., Hagoort, P., Oostenveld, R., & Van Gerven, M. A. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24*(2), 426-434.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience, 18*(1), 555-586.

Small, D. M., Gregory, M. D., Mak, Y. E., Gitelman, D., Mesulam, M. M., & Parrish, T. (2003a). *Dissociation of neural representation of intensity and affective valuation in human gustation*.

Small, D. M., Gregory, M. D., Mak, Y. E., Gitelman, D., Mesulam, M. M., & Parrish, T. (2003b). Dissociation of neural representation of intensity and affective valuation in human gustation. *Neuron, 39*(4), 701-711.

Smith, M. (1987). The Humean theory of motivation. *Mind, 96*(381), 36-61.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences, 11*(01), 1-23.

Sommer, T., Rose, M., Weiller, C., & Büchel, C. (2005). Contributions of occipital, parietal and parahippocampal cortex to encoding of object-location associations. *Neuropsychologia, 43*(5), 732-743.

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron, 35*(6), 1157-1165.

Spreng, R. N., & Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of cognitive neuroscience, 22*(6), 1112-1123.

Sterelny, K. (1990). *The representational theory of mind: An introduction*: Basil Blackwell.

Stich, S. (1999). *Deconstructing the Mind*. New York, New York: Oxford University press.

Swann, N. C., Cai, W., Conner, C. R., Pieters, T. A., Claffey, M. P., George, J. S., . . . Tandon, N. (2012). Roles for the pre-supplementary motor area and the right inferior frontal gyrus in stopping action: electrophysiological responses and functional and structural connectivity. *Neuroimage, 59*(3), 2860-2870.

Swick, D., Ashley, V., & Turken, U. (2008). Left inferior frontal gyrus is critical for response inhibition. *BMC neuroscience, 9*(1), 102.

Thorpe, S., Rolls, E., & Maddison, S. (1983). The orbitofrontal cortex: neuronal activity in the behaving monkey. *Experimental Brain Research, 49*(1), 93-115.

Thut, G., Schultz, W., Roelcke, U., Nienhusmeier, M., Missimer, J., Maguire, R. P., & Leenders, K. L. (1997). Activation of the human brain by monetary reward. *Neuroreport, 8*(5), 1225-1228.

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*(5811), 515-518.

Tremblay, L., & Schultz, W. (2000). Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *Journal of neurophysiology, 83*(4), 1864-1876.
Ungerleider, L. G., & Haxby, J. V. (1994). 'What'and 'where'in the human brain. *Current opinion in neurobiology, 4*(2), 157-165.
Van Den Heuvel, M. P., & Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *European neuropsychopharmacology, 20*(8), 519-534.
van Oers, C. A., Vink, M., van Zandvoort, M. J., van der Worp, H. B., de Haan, E. H., Kappelle, L. J., . . . Dijkhuizen, R. M. (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*(1), 885-893.
Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in cognitive sciences, 12*(11), 418-424.
Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory. *Cognitive, Affective, & Behavioral Neuroscience, 3*(4), 255-274.
Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature, 382*(6592), 629.
Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W. F., & Heiss, W. D. (2005). Role of the contralateral inferior frontal gyrus in recovery of language function in poststroke aphasia. *Stroke, 36*(8), 1759-1763.
Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W. F., & Heiss, W. D. (2007). The right inferior frontal gyrus and poststroke aphasia. *Stroke, 38*(4), 1286-1292.
Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive coding of task-relevant information in human frontoparietal cortex. *Journal of Neuroscience, 31*(41), 14592-14599.
Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D. F., & Büchel, C. (2006). Dissociable systems for gain-and loss-related value predictions and errors of prediction in the human brain. *Journal of Neuroscience, 26*(37), 9530-9537.
Yacubian, J., Sommer, T., Schroeder, K., Gläscher, J., Braus, D. F., & Büchel, C. (2007). Subregions of the ventral striatum show preferential coding of reward magnitude and probability. *Neuroimage, 38*(3), 557-563.
Zaehle, T., Geiser, E., Alter, K., Jancke, L., & Meyer, M. (2008). Segmental processing in the human auditory dorsal stream. *Brain research, 1220*, 179-190.
Zald, D. H., & Pardo, J. V. (1997). Emotion, olfaction, and the human amygdala: amygdala activation during aversive olfactory stimulation. *Proceedings of the National Academy of Sciences, 94*(8), 4119-4124.
Zeki, S., Watson, J., Lueck, C., Friston, K. J., Kennard, C., & Frackowiak, R. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience, 11*(3), 641-649.

Appendix

Image Acquisition

Neuroimaging was performed using a Seimens Prisma 3.0T scanner with a 32-channel head coil at the Harvard Brain Sciences Center in Cambridge, MA. A high-resolution scan was collected prior to functional data acquisition. The echo-planar imaging (EPI) pulse sequence for functional scans used a 2000ms TR with 190TRs per functional run. Stimuli were presented using Psychtoolbox software for Matlab.

Quality Control

All subjects' data was screened for excessive motion. Subjects were excluded from data analysis who were two standard deviations from the group mean on two of the three following motion parameters: average absolute motion per run, average number of movements greater than .5mm per run, and average per run signal to noise ratio. Five of 40 subjects were excluded for excessive motion. One subject was excluded due to technical issues with stimuli presentation during scanning. Three subjects were excluded for getting fewer than 70% of the task attention check questions correct. Finally, for one subject, only 11 out of 12 runs were included in analyses due to technical issues during scanning.

Preprocessing

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 5 TRs were removed from each run. After performing despiking and slice time correction, each subject's EPI images were spatially registered to the first volume of the second run using cubic polynomial interpolation. Motion parameters and temporal trends were removed from the data used for the multivariate analyses (this was performed as part of the HRF deconvolution in the univariate analyses). Data used in the univariate analyses were smoothed with a Gaussian kernel at 6.6mm FWHM (the equivalent of 3 voxels). Data used in the multivariate analyses were left unsmoothed. The mean signal level at each TR was regressed out of the data used for the functional connectivity analyses only.

Whole-Brain Searchlight Analysis

We used a whole-brain searchlight procedure (Kriegeskorte & Bandettini, 2007) with all of our multi-voxel pattern analyses. Following the approach of Mitchell et al. (2008), we created a single image for each trial by averaging over the temporal interval from 6 to 14 seconds after the start of the 8-second delay period to account for hemodynamic lag. We conducted our searchlight analyses using the Searchmight Toolbox (Pereira & Botvinick, 2010). A cube with a 2-voxel (6.6mm) radius was centered at each voxel and a Gaussian naïve Bayes classifier was used to probe the surrounding region for proposition content. Non-edge neighborhoods contained 124 voxels.

Full Results

Regions exhibiting greater activity for winning desire trials > correct belief or losing desire trials during reveal arrow presentation

Lateralization	Region*	TLRC coordinates			Peak t-score	Cluster Size
		x	*y*	*z*		
Winning desire > correct belief						
B	SFG	+/-2	-53	32	6.21	28782
R	MTG	-62	28	-5	5.18	2328
L	Del	9	77	-19	5.17	1509
L	Cer	2	50	-38	5.87	480
L	PG	46	-5	34	5.26	306
L	Fus	48	50	-10	5.18	171
Winning desire > losing desire						
R	MFG	-4	-47	43	3.69	12574
R	IPL	-46	50	50	4.17	481
R	ITG	-55	52	-14	4.83	477
L	MTG	55	50	-5	4.91	476
L	IPL	53	37	47	5.16	441
R	MFG	-26	-18	56	4.64	276
L	Ang	42	72	30	4.45	172
R	Ang	-46	68	32	4.48	150
L	Cer	35	59	-32	3.79	137
Conjunction						
B	MFG	0	48	16	n/a	13143
R	SMG	-46	41	34	n/a	477
R	ITG	-59	30	-19	n/a	414
L	IPL	33	46	36	n/a	318
R	SFG	-20	-18	43	n/a	282
L	IFG	42	-3	23	n/a	265
L	Fus	46	48	-16	n/a	162
L	Cer	33	55	-32	n/a	141
L	MFG	20	-31	36	n/a	95
R	MTG	-48	68	21	n/a	38

Ang, angular gyrus; Cer, cerebellum; Del, delclive; Fus, fusiform gyrus; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; MFG, medial frontal gyrus; MTG, medial temporal gyrus; PG, precentral gyrus; SFG, superior frontal gyrus; SMG, supramarginal gyrus. *Indicates anatomical region containing voxels of peak activation for entire cluster although in many cases cluster extends through multiple regions. For conjunction, indicates region in the approximate center of the cluster.

Regions exhibiting greater mean levels of activity for desire > belief/think trials

Lateralization	Region*	TLRC coordinates			Peak t-score	Cluster Size
		x	*y*	*z*		
Desire > belief						
B	MidFG	+/-31	-53	10	4.07	2453
L	SPL	51	61	34	6.96	789
L	dlPFC	42	-14	45	5.97	304
L	MTG	59	33	-1	3.78	252
Desire > think						
L	SPL	35	66	47	3.92	895
L	MidFG	31	-51	10	5.32	403
L	SFG	20	-38	41	4.23	339
L	MTG	58	35	1	3.85	177
L	dlPFC	46	-14	39	4.12	156
R	IPL	-40	66	41	3.81	141
Conjunction						
L	pIPL	51	52	21	n/a	628
L	MidFG	15	-44	-5	n/a	280
L	MTG	57	30	-12	n/a	153
L	MidFG	9	-40	30	n/a	153
L	dlPFC	44	18	36	n/a	111
L	MidFG	7	31	41	n/a	38
L	SFG	11	-20	54	n/a	21

dlPFC, dorsolateral prefrontal cortex; IPL, inferior parietal lobule; MidFG, middle frontal gyrus; MTG, medial temporal gyrus; pIPL, posterior inferior parietal lobule; SFG, superior frontal gyrus; SPL, superior parietal lobule. *Indicates anatomical region containing voxels of peak activation for entire cluster although in many cases cluster extends through multiple regions. For conjunction, indicates region in the approximate center of the cluster.

Regions exhibiting greater mean levels of activity for belief > desire/think trials

Lateralization	Region*	TLRC coordinates			Peak t-score	Cluster Size
		x	*y*	*z*		
Belief > desire						
R	SPL	-9	63	61	4.4	898
R	Prec	-13	81	43	3.89	702
R	PC	-9	57	10	5.33	314
R	PHG	24	37	-8	3.8	162
L	PC	9	55	8	4.28	131
Belief > think						
R	SPL	-11	68	58	4.22	224
R	IPL	-37	44	41	3.76	136
Conjunction						
R	Prec	-20	63	43	n/a	160

IPL, inferior parietal lobule; PC, posterior cingulate; PHG, parahippocampal gyrus; Prec, precuneus; SPL, superior parietal lobule. *Indicates anatomical region containing voxels of peak activation for entire cluster although in many cases cluster extends through multiple regions. For conjunction, indicates region in the approximate center of the cluster.

Peak activation regions for non-belief related univariate contrasts

Lateralization	Region*	TLRC coordinates			Peak t-score	Cluster Size
		x	*y*	*z*		
Think > belief						
R	IFG	-51	-25	1	4.06	50
Desire > belief						
B	MFG	+/-31	-53	10	4.01	2453
L	SPL	51	61	34	6.96	789
L	dlPFC	42	-14	45	5.97	304
L	MidTG	59	33	-1	3.78	252
Conjunction						
R	IFG	-46	-27	-10	n/a	24

dlPFC, dorsolateral prefrontal cortex; IFG, inferior frontal gyrus; MFG, medial frontal gyrus; MidTG, middle temporal gyrus; SPL, superior parietal lobule. *Indicates anatomical region containing voxels of peak activation for entire cluster although in many cases cluster extends through multiple regions. For conjunction, indicates region in the approximate center of the cluster.

Regions of above chance location classification

Lat	Region*	TLRC Coordinates			Peak Accuracy-Chance	Peak t-score	Cluster Size
		x	*y*	*z*			
B	Cun	+/-2	83	10	18.64	6.04	10619

Cun, cuneus. *Indicates anatomical region containing voxels of peak classifier accuracy although in many cases cluster extends through multiple regions.

Regions of above chance object classification

Lat	Region*	TLRC Coordinates			Peak Accuracy-Chance	Peak t-score	Cluster Size
		x	*y*	*z*			
Object across attitude							
R	STG	-42	-7	21	3.5	4.93	3555
L	Cer	2	61	-36	3.63	5.31	3372
L	Ins	42	-7	-1	3.41	4.9	1567
R	Cul	-44	44	27	3.65	6.33	1253
L	Prec	4	59	54	3.07	5.56	403
R	PG	-26	22	47	3.3	4.9	393
L	SPL	35	61	56	3.18	4.08	334
L	Cing	13	26	28	2.85	4.1	334
Conjunction of object classification with single test attitudes							
L	PHG	18	28	-21	n/a	n/a	168
L	Cer	4	68	-27	n/a	n/a	35
L	Fus	44	61	-16	n/a	n/a	20

Cer, cerebellum; Cing, cingulate; Cul, culmen; Fus, fusiform gyrus; Ins, insula; Prec, precuneus; PG, precentral gyrus; PHG, parahippocampal gyrus; SPL, superior parietal lobule; STG, superior temporal gyrus. *Indicates anatomical region containing voxels of peak classifier accuracy although in many cases cluster extends through multiple regions.

Regions of significantly better object classification in desire or belief trials relative to think trials

Lat	Region*	TLRC Coordinates			Peak Accuracy Difference	Peak t-score	Cluster Size
		x	*y*	*z*			
Object in desire – object in think							
L	PL	7	35	67	7.5	4.29	342
L	Cun	18	70	30	7.86	4.81	238
L	Put	22	-9	6	7.71	4.99	224
L	Cing	15	-16	28	7.05	4.69	208
Object in belief – object in think							
L	PC	0	63	10	9.21	5.11	291
R	pSTG	-42	30	14	8.33	5.09	271
R	PL	-9	15	45	8.5	5.71	197

Cing, cingulate; Cun, cuneus; PC, posterior cingulate; PL, paracentral lobule; Put, putamen; pSTG, posterior superior temporal gyrus. *Indicates anatomical region containing voxels of peak classifier accuracy although in many cases cluster extends through multiple regions.