SHARING ALL THE WAY TO THE BANK: A NEUROIMAGING INVESTIGATION OF DISCLOSURE, REWARD, AND SELF

by

WILLIAM EVERETT MOORE III

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2015

DISSERTATION APPROVAL PAGE

Student: William Everett Moore III

Title: Sharing All the Way to the Bank: A Neuroimaging Investigation of Disclosure, Reward, and Self

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Jennifer H. Pfeifer	Chairperson
Elliot T. Berkman	Core Member
Nicholas B. Allen	Core Member
Mark Alfonso	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2015

DISSERTATION ABSTRACT

William Everett Moore III

Doctor of Philosophy

Department of Psychology

September 2015

Title: Sharing All the Way to the Bank: A Neuroimaging Investigation of Disclosure, Reward, and Self

No neuroimaging investigation to date has considered the effects of social context on self-referential processing, despite the fact that the hypothesis that people engage different selves in different contexts has been with psychology for more than a century. To address this gap in the empirical record, a suite of three functional magnetic resonance imaging (fMRI) experiments was conducted in order to assess patterns of neural activity associated with self-referential (compared to non-self-referential) processes (Experiment 1), computational models of reinforcement-learning processes (Experiment 2), and social context modulation of personally relevant cognition (Experiment 3). I demonstrate that distinct patterns of neural activity in cortical midline structures and the mesial ventral striatum are associated with thinking about the self privately, sharing information about the self with a parent, and sharing with a friend. These differentiated disclosure responses (Experiment 3) are evident at the whole brain level and in regions of interest defined by functional activity in independent tasks of self (Experiment 1) and reward (Experiment 2). In addition to providing empirical evidence for contextually differentiated self-representations in the brain, this dissertation validates the use of fMRI paradigms designed to functionally localize self-referential and reward-related activity either independently or in conjunction, as well as distinguish components of ventral striatal activity unique to each task. Finally, I consider strategies for approaching future investigations of self and social cognition in terms of reinforcement learning.

CURRICULUM VITAE

NAME OF AUTHOR: William Everett Moore III

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of North Carolina at Chapel Hill

DEGREES AWARDED:

Doctor of Philosophy, 2015, University of Oregon
Master of Science, 2011, University of Oregon
Bachelor of Science, 2006, University of North Carolina at Chapel Hill

AREAS OF SPECIAL INTEREST:

Neuroscience
Functional Magnetic Resonance Imaging

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, University of Oregon Psychology Department, 2009-2015

PUBLICATIONS:

Jankowski, K.F., Moore, W.E. III, Merchant, J.S., Kahn, L.E., & Pfeifer, J.H. (2014). But do you think I'm cool? Developmental differences in striatal recruitment during direct and reflected social self-evaluations. *Developmental Cognitive Neuroscience*, *8*, 40-54.

Moore, W.E., Merchant, J.S., Kahn, L.E., & Pfeifer, J.H. (2014). Like me?: ventromedial prefrontal cortex is sensitive to both personal relevance and self-similarity during social comparisons. *Social Cognitive and Affective Neuroscience*, *9*(4), 421-426.

Bruce, J., Fisher, P.A., Graham, A.M., Moore, W.E., Peake, S.J., & Mannering, A.M. (2013). Patterns of brain activation in foster children and

nonmaltreated children during an inhibitory control task. *Development and Psychopathology*, *25*(4), 931-941.

Peake, S.J., Dishion, T.J., Stormshak, E.A., Moore, W.E., & Pfeifer, J.H. (2013). Risk-taking and social exclusion in adolescence: neural mechanisms underlying peer influences on decision making. *NeuroImage*, *82*, 23-34.

Moore, W.E, Pfeifer, J.H., Masten, C.L., Mazziotta, J.C., Iacoboni, M., & Dapretto, M. (2012). Facing puberty: associations between pubertal development and neural responses to affective facial displays. *Social Cognitive and Affective Neuroscience*, *7*(1), 35-43.

Pfeifer, J.H., Moore, W.E., Oswald, T.O., Masten, C.L., Mazziotta, J.C., Iacoboni, M., & Dapretto, M. (2011). Entering adolescence: resistance to peer influence, risky behavior, and neural changes in emotion reactivity. *Neuron*, *69*(5), 1029-1036.

ACKNOWLEDGMENTS

I would like to express my sincerest thanks to my Esteemed Committee Members for their patience, wisdom, and guidance in the preparation of this manuscript. Specifically, I would like to thank: Dr. Nicolas B. Allen for his infectiously positive skepticism, without which my search of the problem space would have been considerably less fruitful and less pleasant. Dr. Mark Alfano, who groks in but an instant ideas I have wrestled for an age, for reminding me that a test is only as good as the questions it informs. Dr. Elliot T. Berkman, whose earnest love of science is as deep as is his profound knowledge of it. I would not be the scientist I am without his tutelage or outstanding examples of optimally resolving curiosity through pragmatic satisficing. Finally, I would like to express my deepest gratitude to Dr. Jennifer H. Pfeifer, whose mentorship and model leave me different than I was. Were it not for her I would know not how to write or write a shell script or a voxel or a parcel of the parcellated mind. I would like to acknowledge her commitment not only to methodological rigor in a field fraught with all the noise of teenagers and time, but to the brain as a dynamic system and to science as an effort that is fundamentally collaborative. I would like to thank Drs. Tor Wager, Russ Poldrack, Tal Yarkoni, Luke Chang, and Diana Tamir for the seminal roles they played in my development as a neuroimager. Scott Wattrous, a technoshaman *par excellence*, taught me how to actively integrate the practice of science into one's heart, mind, and life. I will be forever grateful to have learned at the feet of such an epic master of the magnet. Dr. Jolinda Smith is responsible for my fledgling understanding of the whims of the great magnet, according to which it is said that all things flow, and her repeated, patient explanations lie at the core of what makes me ok with kspace. Chuck Theobald's resolute commitment to information betrays a passion for text and a wisdom to maintain it I hope to carry with me always, and I much appreciate his transmission of said signals. David Anderson, an arcane sorcerer of the hermetic tradition, has given me perspective and insight into the mind of a true scientist, with a relentless commitment to exploring the universe. Ryan Giuliano pushes me always to grow from our collective failure to ignore the white

bear, and I am honored to have known a sorcerer whose school defies classification. I feel fortunate to have likewise stood and studied with the warlock Brian Clark, who not only encourages me to be as skeptical of empirical conclusions as moral ones, but enriches my understanding of the difference. Shannon Peake is the finest office mate one ever might know, and I will always be grateful for the kindness, wisdom, and insight with which he is so generous. LEK's harmonies transcend the acoustic and evince themselves in raw information, and I will forever be in awe of her ability to juggle in literal or metaphorical senses. I would like to thank Junaid Merchant for turning me on to Social Neuroscience, for sharing with me his brilliant ideas about it, for his epic patience with complaints about said science, and for decades of lessons concerning the former aspect of our discipline. I would like to thank Morgan Johnson, premier nerd wrangler, for helping me to rediscover a part of myself I thought long dead, for introducing me to brilliant folk from all walks of life, and for prospectively agreeing to teach me how to surf. I would like to thank, Tanner Bower, for helping me to discover a part of myself I only ever suspected was alive, and his mentorship, kindness, and wisdom enrich my science and my life. The Upstanding Gentlemen Next Door provided the live soundtrack to which the bulk of this work was composed, and were their jams less icy, the process would have been far less pleasant. Nathan Alter, in particular, deserves thanks for reminding me that writing is a broader pursuit than that of scientific publication. I would like to thank my parents, Bill and Nancy Moore, whose patience with respect to all my self-disclosures continues to prove invaluable. I would never have succeeded at much without their constant love and support. I would like also to thank my Aunts Dr. Allison O'Brien and Dr. Carol-Anne Coyle, for their patience, guidance, and perspective in completing this challenging process. I would like to thank Jake, Wes, and Mona Hartman for their tolerance and support throughout the completion of my degree, but more importantly for making me a part of their family and for sharing with me Alexandra, my role model and muse. It is indeed my most peculiar self that resides in her mind, or hers in mine, and I cannot imagine my life without our potential social we.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I
BACKGROUND

Self-relevance as Value-based Decision Making

> In each kind of self, material, social, and spiritual, men distinguish between the immediate and actual, and the remote and potential, between the narrower and the wider view, to the detriment of the former and advantage of the latter. One must forego a present bodily enjoyment for the sake of one's general health; one must abandon the dollar in the hand for the sake of the hundred dollars to come. One must make an enemy of his present interlocutor if thereby one makes friends of a more valued circle; one must go without learning and grace, and wit, the better to compass one's soul's salvation. *Of all these wider, more potential selves, the potential social Me is the most interesting.* (James, 1890, p. 191)

William James proposed that distinct selves within an individual are defined by choices between short-term outcomes and long-term consequences. This is, informally, a key difference between *reward* (an immediate consequence) and *value* (a long-term estimate about rewards from now on). The intersection of self, value, and reward are especially interesting in the context of the broad hypothesis that self and reward are fundamentally related processes (Northoff and Hayes, 2011). Relating the interoceptive signals that describe an organism's internal state to the external sensory stimuli in the environment has been separately proposed as an essential function of both reward (Montague and Berns, 2002) and the self (Enzi, et al., 2009). Explaining how people override the temptation to choose a cool, crisp dollar bill for the self of right this second instead of one hundred dollars for some abstract future self is the exact sort of problem that contemporary neuroeconomics approaches attempt to solve by considering the underlying computations (Rangel, Camerer, and Montague, 2008). Coupled with demonstrations that different aspects of the self are differentially valued in the brain (D'Argembeau et al., 2011) and the same neural mechanisms compute social and monetary value (Izuma, Saito, and Sadato,

2008), a logical next step is to extend these findings by demonstrating that, in behavior and in the brain, potential social selves are more highly valued than the immediate private self. This dissertation first presents historical and scientific context for contemporary neuroimaging investigations of the self. Subsequently, empirical approaches to reward are evaluated in terms of their applicability to probabilistic decision-making. Finally, neuroeconomics is considered as a possible avenue from which to mutually inform investigations of the self and reward.

Neural Substrates of Self-relevant Processes

The earliest fMRI investigation of the self was an attempt to resolve a dispute about the underlying causes of the "self-reference memory effect," which describes people's tendency to exhibit superior recall for trait adjectives encoded in terms of the self. One interpretation framed the effect as a logical extension of the depth-of-processing effect, explaining enhanced recall performance as a consequence of the vast amount of self-relevant information in memory, which increases the likelihood that a stimulus presented in this context will be more richly encoded (Klein and Loftus, 1988). The competing hypothesis maintained that self is a special construct which receives privileged information processing status (Rogers et al., 1977). An experiment using fMRI supported the latter hypothesis, demonstrating that trait adjectives encoded in terms of the self were more strongly associated with activity in the ventromedial prefrontal cortex (vmPFC), and that this activity subsequently predicted recall during a surprise memory test (Kelley, et al., 2002). This finding represented a powerful proof-of-concept for the utility of fMRI for addressing questions that cannot be answered with a purely behavioral approach. In addition to a succinct methodological demonstration, it provided strong evidence to directly resolve active scientific debate, as well as the first support for a biological substrate of self-referential processing.

Collectively, these factors led to a spike in neuroimaging investigations of the self, and the tide of findings broadly implicating medial prefrontal cortex (mPFC) in self-relevant and social cognition swelled to critical mass for meta-analytic approaches in just a few short years, linking activity in mPFC across dozens of investigations to making judgments about the enduring characteristics of others (Van Overwalle, 2009), to explicit self-reflection (Van Der Meer, Costafreda, Aleman, and David, 2010), and to processing social stimuli with a high degree of self-relevance (Enzi et al., 2009). The vmPFC in particular exhibits heightened responses during non-comparative judgments about self-similar (as opposed to dissimilar), unfamiliar social targets (Mitchell, Macrae, and Banaji, 2006). However, the vmPFC has alternatively been implicated in representing the degree of personal closeness for social stimuli, responding preferentially to targets with heightened social relevance, regardless of the degree of similarity between the self and judgment target (Krienen, Tu, and Buckner, 2010). A previous attempt in our laboratory to investigate the self along multiple dimensions employed comparative social judgments between personally relevant others and/or the self, providing neuroimaging evidence for the hypothesis that comparisons between the self and similar others preferentially engage a subregion of mPFC, perigenual anterior cingulate cortex (pgACC; Moore, Merchant, Kahn, and Pfeifer, 2014). Additional results characterized the pgACC as simultaneously sensitive to the degree of self-relevance (operationalized here as personal involvement) of a judgment and the extent to which non-self judgment targets were regarded as personally similar.

Across many of these neuroimaging investigations of self-relevant cognition, positive differences in BOLD signal across conditions were actually relative differences, and reflected “less deactivation” rather than “activation” compared to a resting baseline, similar to patterns typically observed in the brain’s Default Mode Network (DMN; Buckner and Caroll, 2007). The DMN is an interconnected group of regions that responds to most stimuli with a marked disengagement from the high levels of coordinated activity observed during stimulus-independent thought. However, a wide variety of tasks assessing self-

and social cognition have been associated with “less negative,” or even explicitly positive, changes in DMN from resting state levels of activity, (D’Argembeau et al., 2005). An analysis of resting state activity and cognitively demanding tasks of episodic retrieval and prospective memory (e.g., imagining the self in the future) suggested that *prospective self-referential cognition* most strongly resembles the DMN at rest, which recalling the past elicits the typical “deactivation” (Whitfield-Gabrieli et al., 2011).

An experiment at the single cell level in deep brain stimulation patients paints a similar picture in the subcallosal cingulate (scCC; an area slightly inferior and posterior to pgACC), suggesting that the individual firing rates of cingulate neurons do not change from a resting baseline in response to the presentation of one’s own name, but that presentation of another person’s name elicits a dramatic increase in the spiking output of these cells (Lipsman et al., 2014). However, this experiment did not vary the personal relevance of “other” names presented, and while findings from single neurons and group-level whole-brain analyses simultaneously implicate mPFC in self-relevant processes, the way that this part of the brain distinguishes between the self and personally relevant others remains unclear. The same investigators who initially mapped the self-reference memory effect onto activity in mPFC have proposed that the general role of this region in self-referential processing may be (1) an extension of a more general system for social cognition, (2) a meta-executive function integral to the coordination of complex attentional states, or (3) a central hub for integrating external sensory cues and interoceptive signals with abstract cognitive and affective states (Moran, Kelley, and Heatherton, 2013). It is this final explanation, the authors maintain, that seems most consistent with our understanding of DMN coherence during stimulus independent thought that is similar for tasks of self-reference but sharply reduced for stimuli that are not particularly engaging on a social cognitive level.

In Pursuit of Reward

Historically, the psychological study of reward has been dominated by reward-learning, as “learning” is far more methodologically convenient to measure than other phenomena that fall under the broad reward umbrella -- e.g. “reward wanting” (i.e. motivation), or “reward liking” (hedonic state; Montague 2007). Unlike pleasure or desire, learning can be readily quantified in terms of behavioral change. Affect and motivation, especially in animal models, are not easy to measure. It is perhaps for these reasons that perspectives on reward have been framed primarily in terms of positive reinforcement for appetitive stimuli that most effectively shape behavior (e.g., fruit juice) (Montague, 2006) – for which there is no negative equivalent (e.g., a “not-juice” stimulus, in the absence of a relative comparison to juice). Although these two trends (reward as positive reinforcer and the dominance of reward-learning paradigms) continue to impact contemporary perspectives and nomenclature on reward, recent advances in neuroimaging and computer science hold considerable promise for understanding the more abstract aspects of reward-related processes.

Learning models: Making the distinction between value and reward

Formal models of animal behavior have been able to account for straightforward kinds of learning since the seminal work of Bush and Mosteller (1951), but Rescorla and Wagner (1972) extended this model to include a parameter that quantifies the abstract associative weight (V) between multiple sensory properties of a stimulus (rather than using the raw probability that the sensory events happened at the same time). This critical addition enabled researchers to account for previously unsolvable problems in classical (Pavlovian) conditioning. The Rescorla-Wagner (R-W) model was not only instrumental in shaping the application of formal reinforcement-learning to more complex kinds of behavior like operant conditioning, but was also seminal in popularizing the use of computational techniques to create normative frameworks for optimal behavior in psychology and neuroscience (Glimcher, Dorris, and

Bayer, 2005). Granted, application of an R-W model to describe a single trial of conditioning based purely on the physical features of stimuli, or even an extension to more complex kinds of learning through incremental reinforcement, seems more than a little remote from value-based decision making in humans. The reason the R-W model continues to be relevant today is because the associative weight, V, can be defined in terms of *any* abstract association, it is not restricted to paradigms with passive stimulus exposure.

Considering a much more complicated experimental situation than single-trial conditioning (but still a relatively simple one), assume that a juice-loving animal is presented with a visual stimulus that elicits a left or right button press. Immediately thereafter, some amount of juice is delivered to him based on predetermined stimulus-response contingencies. By allowing the learning rate to range continuously from 0-1 and using multiple stimulus-response contingencies in the imagined experiment, we can obtain precise estimates of reward (a raw numerical index of “amount of juice right now”) and expected value (an estimate about “how much juice in the future” from each available stimulus-response contingency (i.e., that stimulus and a left or right response) on each trial. These quantities can then be combined to index the extent to which the organism’s predictions deviate from its perceived outcomes. This difference between an organism’s state and its expectations about that state in light of a potentially informative stimulus is known as *prediction error* (PE). When the reward an organism receives from a stimulus perfectly matches the value it expects, there is no prediction error, and no learning takes place. This is called a “fully predicted” reward (Kamin, 1969).

To understand why prediction error is so crucial to reward, consider a slot machine that is completely predictable. If the outcome of paying a quarter and pulling the lever on any given trial was always known, then playing the slots would feel more like feeding a meter where you park bad decisions and less like gambling. There would be no anticipatory thrill at the unlikely (but non-zero) prospect that you hit the jackpot. There would still be would be a reward in the

formal, algorithmic sense, but because the reward has been perfectly predicted, it has lost the motivational component that “keeps us coming back for more” (Schultz, Dayan, and Montague, 1997).

Measuring prediction error signals in the brain

The spiking profile of dopamine neurons that terminate in the nucleus accumbens (nAcc) is best characterized in terms of a “prediction error signal” (Schultz, Dayan, and Montague, 1997). There are, broadly-speaking, two modes of activity that mesolimbic dopamine neurons use to convey prediction error (Fox, et al., 2004; O’Doherty, Hampton, and Kim, 2007). Tonic activity refers to action potentials that occur at low frequency, regular intervals while an organism is at rest or performing a task. Phasic activity refers to concentrated bursts of high frequency spikes that do not occur at regular intervals. When an organism receives an unanticipated positive reward, or a reward whose magnitude is far greater than expected -- in sum, any event for which there is a positive prediction error -- dopamine neurons in the ventral tegmental area (VTA) exhibit phasic bursts of spike output, enervating the nAcc with dopamine. These neurons do not fire action potentials, however, in response to fully predicted rewards (even if the magnitude of that reward is very high). In contexts where an organism has a highly accurate estimate for the EV of its actions and decisions, there is not likely be any PE, and thus little phasic activity in dopaminergic reward circuits. Conversely, when an organism receives a reward with less actual value than was expected, the tonic activity of nAcc dopamine neurons ceases. Thus, positive PE is indexed by phasic bursts of dopamine transmission, while negative PE is tracked by inhibition of tonic dopaminergic activity.

Using model-based fMRI, researchers have been able to describe the BOLD signal with functionally identical computational models to those used to characterize mesolimbic dopamine neurons (Fox, et al., 2004; O’Doherty, Hampton, and Kim, 2007). Although a conservative approach suggests that we can only localize the basal forebrain component of this activity to the “ventral

striatum" with BOLD fMRI, the liberality of the record speaks volumes, as a Google Scholar search for "nucleus accumbens fMRI" returns over 20,000 hits as of May 10, 2014. Issues of nomenclature aside, it is now a well replicated finding that brain's prediction error signal can be measured in the human vS and medial prefrontal cortex, as well as mesencephalic dopamine neurons. In fact, neural correlates of the prediction error signal have more recently been identified all over the brain – not just in tasks of reward-learning, but also in perceptual and attentional processes as well (den Ouden, Kok, and deLange, 2012).

Neuroeconomics: A New Science of Decision Making

Convergent findings in systems neuroscience and behavioral economics have prompted leading scholars in their respective fields to claim that the relatively young field of "neuroeconomics" represents a consilience of decision making and the brain (Glimcher and Rustichini, 2004). While calling it a unified theory of human behavior is, perhaps, overstating the case, neuroeconomics has dramatically strengthened both conceptual and evidence-based bridges between brain and behavior. This interdisciplinary approach allows neuroscientists to constrain their noisy data with formally defined models, while providing economists with the ability to validate and refine said models through empirical testing - not only at the microeconomic, but at the microscopic level. As its name might suggest, neuroeconomics is grounded in a few key principles inherited from the disciplines of neuroscience and economics (summarized below, as informed by Montague and Berns, 2002, among others):

1) Mobile organisms have limited resources in terms of time and energy.
2) Investment of these resources is based on predictions about the consequences for an organism associated with any stimulus or behavior.
3) To assess outcomes and update predictions, an organism needs a common scale that represents value across incomparable domains of information.

4) Normative models of value-based decision-making can be used to test formal predictions about value assignment across multiple levels of analysis.

The first principle is difficult to contest. Human organisms require nutrients in order to stay alive and mates in order to produce viable offspring. Rational decisions about how to deploy a limited supply of time and energy are based on predictions of future outcomes derived from prior experience. While an organism might make an irrational prediction or randomly associate unrelated outcomes with potential actions, a systematic approach for generating reliable expectations about the universe based on the outcomes of previous decisions is central to adaptive fitness. The assumption that behavioral decisions necessitate a common scale for value appears more controversial, but the mere selection of any one response compared to another implies a subjective preference for the option that was chosen. This so-called “revealed preference” indicates that the organism regards its choice as subjectively “better” than the alternative(s). While a random or irrational choice in isolation may thus convey an erroneous preference, a real preference will be evident in the bias exerted over repeated decision-making. It is likely that multiple neural systems compute value at different levels of stimulus complexity (Rangel, Camerer, and Montague, 2008), but all value-based choices must rely on an ultimately common neural scale. This scale, or currency, is required to synthesize information across qualitatively distinct domains of interoceptive and sensory information (Schultz, Dayan, and Montague, 1997). The final principle is more akin to a data-driven observation than a fundamental assumption, but it is the element most central to the conscilience of brain and decision touted by advocates of the neuroeconomics approach (Glimcher and Rustichini, 2004). The introduction of normative frameworks allows neuroscientists to simultaneously constrain noisy data with reinforcement learning algorithms that describe optimal behavior for decision problems that have no known ideal solution. The crucial component of the class of models known as prediction valuation models is the reward prediction error signal, which indexes the discrepancy between actual rewards and expectations

about reward outcomes. Rangel, Camerer, and Montague (2008) have presented a framework for the neurobiology of value based decision making that describes the basic components of the decision process (**Figure 1**), and proposes three distinct valuation systems to account for different kinds of value assignment (**Figure 2**). One strength of this framework is that the basic computations are described in terms of psychological phenomena, but are also organized so as to be readily computable through a variety of models, providing researchers with the tools to model brain and behavior at multiple levels of analysis.

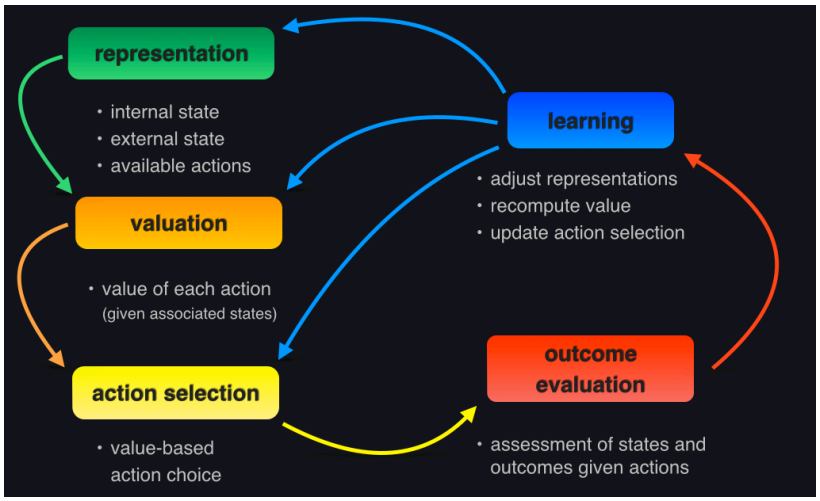

Figure 1. Biological computations underlying value-based decision making. This diagram describes the information processes underlying any choice, and it can be applied to describe observed behavior and neural activity in the formal terms of a broad class of predictor-valuation models. Adapted from Rangel et al. (2008).

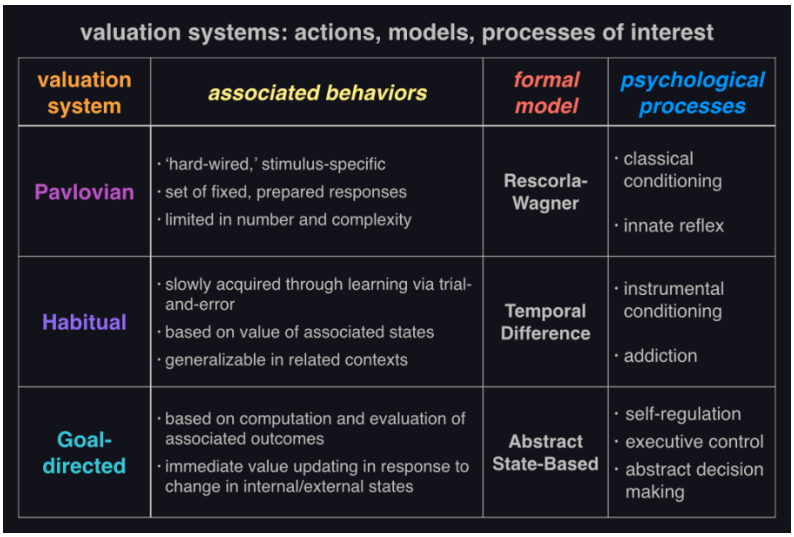

Figure 2. Valuation systems for brain and behavior. This chart describes differences in the behavioral, formal, and psychological phenomena best described by each of the Pavlovian, Habitual, and Goal-directed valuation systems, as proposed in Rangel et al. (2008).

Contemporary reinforcement-learning systems descended from the machine learning literature involve three formal components: a reward function, a value function, and a "policy" (Montague and Berns, 2002). An experiment presents decision-making organisms (or "agents") with a finite number of possible "states," which is limited by the available combinations of stimulus-response contingencies. Each state is associated with a reward – a scalar quantity that indexes "how good" the current state is for the organism, right now (which can be negative in the event that the state is "not good"). The organism can select actions (e.g. stimulus-relevant motor output) to transition from one state to another. Here, reward is likened to a context-dependent assessment of immediate consequences for an organism, whereas value describes a more general, long-term assessment of prospects for the future from the current state in light of all subsequent states and domain-free estimates of the associated rewards (Montague, 2007). The value function yields an estimate that indexes "how good" a state is for the organism, considering rewards associated with the immediate state as well as possible future states – formally equating to the *total reward* an agent can expect from the current state, and all subsequent states in

the future. The role of the policy, in turn, is to formally relate states to actions, defining the probability that any action results in a subsequent state.

CHAPTER II
MOTIVATION FOR THE CURRENT STUDY

Value as the Link between Self and Reward

The main difference between value-based decision making frameworks for the study of reward-related processes from the approach of the animal behavior tradition is a central focus on value, rather than reward. The experience of reward occurs in an immediate temporal context. Value, on the other hand, is defined by expectations an organism has about consequences that have not yet transpired. For this reason, *the assignment of value depends on an organism's ability to represent itself in the future*, and thus we return to the self.

As summarized in Chapter I, activity in cortical midline structures has been repeatedly implicated in self-evaluation across multiple domains (e.g., relevance, similarity, closeness). The literature consistently reflects the involvement of mPFC and pgACC in self-referential processes, and it is relatively safe to assume that cortical midline structures are involved in psychological processes of self at the implementation level. Most often, the paradigms employed would suggest that the problem being solved at the process level is a binary "me" versus "not me" judgment (although it should be noted that the actual information processing problem and the ostensible "problem" suggested by task instructions may widely diverge).

Reports that activity in mPFC differentiates self and close others (e.g., Heatherton et al., 2006) imply that a straightforward representation of "internal to the organism" vs. "external to the organism" might biologically differentiate self from non-self. However, mPFC also computes qualitatively different kinds information about the self (e.g., Krienen et al., 2010, Moore et al., 2014), which suggests that an organismal judgment of internal/external may be insufficient for abstract self-evaluation that integrates information across distinct domains (e.g.,

group identity, self-relevance, self-similarity). While it is certainly possible that such an algorithm might ultimately return a simple “me” versus “not me” distinction, it seems more likely that the richly elaborated nature of self-referential stimuli depends on fine-grained computational processes.

Compelling empirical evidence for the relationship between self and reward comes from a suite of experiments centered on the intrinsic reward associated with self-disclosure (Tamir and Mitchell, 2012). Using fMRI, the authors showed that self-referential stimuli elicited stronger responses in the vS, vmPFC, and ventral tegmental area (VTA) compared to other-specific or non-social stimuli, indicating that self-disclosure is associated with activity in the neural structures associated with value computation (Batra, Kable, and Glimcher 2013) This finding was complemented by behavioral experiments showing that participants were willing to forgo monetary rewards in order to share information about themselves, based on the computation of a point of subjective equivalence (PSE) or dollar amount at which participants would electively self-disclose despite financial loss. Together, these studies demonstrated that people are motivated to share information about the self with others, and -- given similar patterns of reward-related BOLD signal for self-evaluations versus non-social evaluations -- that self-referential cognition is, itself, rewarding.

The study of reward presents a nearly identical problem of evaluating qualitatively distinct stimuli, internal states, and behaviors, and well-replicated findings point to neurons in ventromedial prefrontal cortex as the biological substrate for the common currency of subjective value (Louie and Glimcher, 2012), a finding reflected in the human neuroimaging literature, albeit with more coarse resolution (Bartra, McGuire, and Kable, 2013). The opinions of others can modulate value computation in both vS and vmPFC, and this influence on subjective value in the brain continues for at least half an hour after exposure to social information (Zaki, Schirmer, and Mitchell, 2011). Subsequent research demonstrated that the computation of value in vmPFC can also be regarded as “person-invariant” (i.e., does not differ after controlling for the relative value

individuals place on their own and others' gains; Zaki, Lopez, and Mitchell, 2014). This means that the common neural currency of subjective value is computed not only independently of reward receipt, but of reward recipient.

The neural responses elicited by tasks that combine elements of self-relevance and value computation look similar to what we would expect, given that reports from independent investigations of either process routinely implicate vS and vmPFC. Since the psychological processes cannot be disentangled, however, it is possible that the observed patterns of activity are entirely driven by self or value. In a paradigm specifically designed to compare processes of self and reward, a conjunction across contrasts indexing high versus low personal relevance and positive versus negative reward outcomes revealed activity in both pgACC and vS, while a direct comparison of self and reward instead elicited activity in the insulae and mid-cingulate (Enzi et al., 2009). A strict neuroeconomist would likely refrain from regarding these results as evidence that common psychological mechanism underlies self-relevance and value computation per se, as no computation of value was not described in terms of a formal model or isolated from other decision components. To infer that one or more common processes underlie self-relevance and reward-related processing more generally, however, is not unreasonable.

In a theoretical perspective on self and reward, Northoff and Hayes (2011) proposed that the assignment of personal relevance is actually a value computation. If a stimulus comes into the sensory awareness of an organism, the assessment of relevance to the self begins well before the stimulus reaches the threshold of conscious awareness, and that it is value computation and the assignment of affect that determine whether or not any stimulus is regarded as personally relevant. Self-referential cognition, however, is not just the logical extension of self-relevance (**Figure 3**).

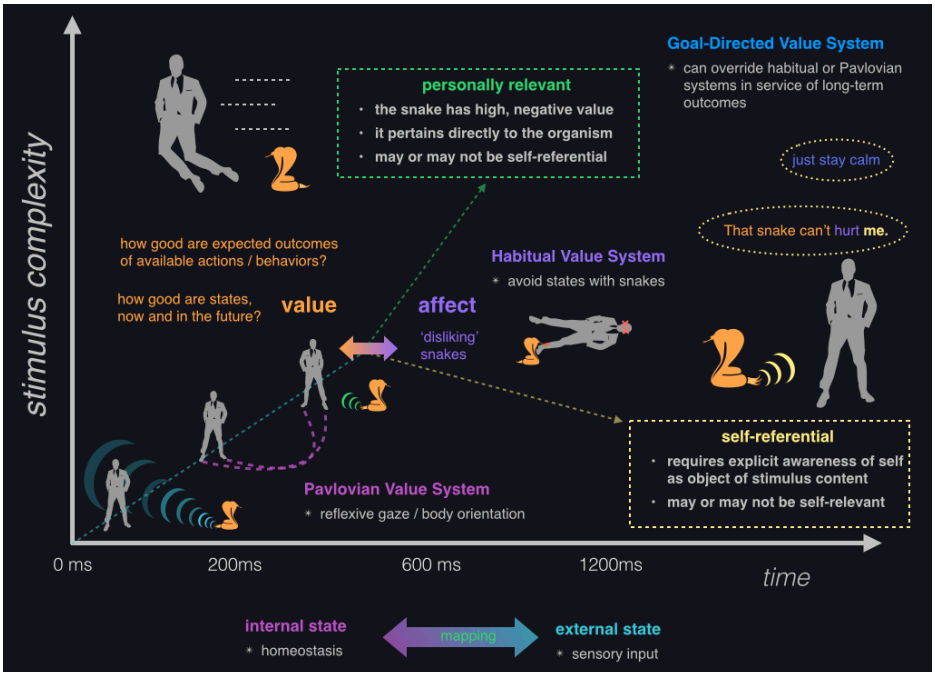

Figure 3. Self-reference, personal relevance, and value assignment. The y-axis represents stimulus complexity, while the x-axis represents time. As a stimulus is perceived initially, Pavlovian reflexes control movement that orients the agent toward the auditory stimulus. As the value and affect associated with the stimulus are compared, an assessment of personal relevance is applied. A stimulus like the snake has high relevance for someone standing behind the snake, and the Habitual Value System most likely has instilled a general disinclination toward places where snakes might be likely to appear. For the snake to be self-referential, it must enter conscious awareness. If the general disinclination toward snakes can be overcome through use of the Goal Directed Value System, the agent can maintain control over an inclination to run away, and instead view the snake as a potential food source.

Although “relevance” is a somewhat arbitrarily selected label for the psychological process driving observed differences in mPFC activity, it is a strategic semantic choice for a process that has been alternatively referred to the literature as distance from the self, personal significance, personal involvement, self-relevance, self-relatedness, or self-specificity (Abraham, 2013). “Relevance” is the most unique and consistently applied term in the class of twists on the theme that generally operationalizing the same concept (although there are occasional inconsistencies), save for one: self-reference. The factor that distinguishes self-reference from self-relevance (or its aliases) is that self-

referential stimuli require explicit awareness of the objective, content-based self. Because all stimuli involve the process-based, subjective self to some extent, we cannot modulate its presence without pharmacological or neural perturbation. We likewise cannot ask people to control an implicit or spontaneous assignment of self-relevance or self-reference judgment, especially for social stimuli (Mussweiler, 2003; Mussweiler, Rüter, and Epstude, 2004). As Abraham (2013) acknowledges, we cannot control implicit aspects of self-relevance or self-reference, but her recommendation that explicitly self-referential cognition be excluded from future investigations does not seem like an ideal way to proceed, given that social stimuli are equally likely to spontaneously evoke implicit reference to the self (Mussweiler, 2003; Mussweiler, Rüter, and Epstude, 2004).

Returning to the opening quote from Chapter I of this dissertation, in addition to these thoughts, William James also claimed that each person has "as many different social selves as there are distinct groups of persons about whose opinion he cares" (p. 294) and that the social self "ranks higher than the material self" (p. 314). He also suggested that the number of social selves an individual has is determined by the assignment of value to the opinions of others, and that various kinds of selves are likewise compared against each other via evaluative judgments. Finally, as noted earlier, James took strongest interest in the content-based aspects of the future social selves we may become (p. 191).

Collectively, these observations from James hint at a novel experimental approach for testing the hypothesis that an individual's multiple selves are distinctly represented in the brain. Through the use of *prospective* social context (i.e., answering binary questions about the self that will later be shared with a friend or parent via email), we can test whether private evaluations of the self are represented differently in the brain from decisions about potential social selves, and whether the prospective contexts of sharing with a friend and sharing with a parent can be likewise differentiated at the neural level. Tamir and Mitchell's (2012) demonstration that self-disclosure is intrinsically rewarding can thus be extended in two ways. First, because prospective disclosures require allows us

asses the long-term value in social sharing, rather than the fleeting, “risky thrill of self-disclosure” (Dahl, 2008). Second, in addition to differentiating self evaluations in a private context from prospectively social ones, comparing disclosures to parents and friends can test the existence of distinct neural representations of multiple social selves as well as the “potential social Me.”

Stated Goals

The overarching goal of this dissertation is to demonstrate the effects of prospective social context on the neural correlates of personal relevance. The first of the three fMRI experiments described here aims to distinguish self-referential from non-self-referential evaluations in cortical midline structures (CMS), a set of regions routinely implicated in stimuli pertaining to the self (Denny et al., 2012). The second experiment is designed to characterize the neural processes associated with probabilistic decision making in terms of a formally defined neural network model. In the second experiment, the reward prediction error signal (the crucial component that drives reinforcement-learning algorithms) is expected to correlated with activity in ventral striatal (vS) BOLD signal, a proposed index of the dopaminergic prediction error signal computed by midbrain dopamine neurons (Schultz, Dayan, and Montague, 1997; Fox et al., 2004). By assessing the results from these first two tasks in conjunction, I aim to demonstrate functional neural overlap in independent tasks of self evaluation and value based decision making -- replicating Enzi and colleagues’ (2009) finding that personal relevance and reward outcome elicit overlapping patterns of BOLD response. I propose to extend their hypothesis by demonstrating that reward prediction error, a more precisely defined and meaningful analog of reward signal in the brain, shares a common neural substrate with personal relevance across independent tasks.

The third experiment will address the primary question of interest: Does prospective social context modulate the neural representation of self? In this experiment, participants performed trivial self-evaluations in either a private

context or a prospectively social context (i.e. with a parent or friend who would be informed of their answers via email) in order to elicit the effects of social influence on self-referential cognition in a physically isolated context. This experiment demonstrates that the neural correlates of self-evaluation differentiate between private and potentially social aspects of the self. In addition to providing the first empirical evidence for potential social selves in the brain, this dissertation provides an empirically tested and meta-analytically validated paradigm for isolating neural activity associated with self-referential cognition.

Predictions

I predict the contrast of self versus change (Experiment 1) will elicit stronger responses in the ventral striatum (vS) and medial prefrontal cortex (mPFC), to include the perigenual aspect of anterior cingulate cortex (pgACC). The extent to which personal relevance can be inferred from this contrast is contingent on the assumption that the control condition, in which participants were explicitly instructed not to answer the question about themselves personally, can be fairly regarded as a subtractive means of indexing personal relevance. This assumption, however, has considerable empirical precedent, and is the basis for much of our collective understanding of self-relevant processing. Additional support for this inference will come in the form of a non-exploratory, formal reverse inference of the contrast, self versus change, in order to index the extent to which the reported neuroimaging literature reflects correlations with results from conceptual similar topics of interest (e.g., “self,” “autobiographical,” “self-referential”).

In Experiment 2, I predict that reward prediction error, as calculated formally in terms of the difference between reward and expected value on each trial, will correlate with BOLD signal in the bilateral vS and vmPFC. BOLD signal in human vS has been repeatedly link mapped to the dopaminergic prediction error signal observed at the single unit level (Fox et al., 2004), and the extent to which behaviorally derived reward prediction error parameter captures the vS and vmPFC responses to reward outcome events should provide sufficient

evidence of neural computation of prediction error signal to substantiate an influence about the brain's prediction error signal. Although the anticipated result has been roundly replicated, and, like Experiment 1, would not represent a novel contribution to the field of social neuroscience taken purely in its own context, the statistical results concerning reward prediction error and personal relevance will be conjointly tested to assess the neural overlap of self and reward.

I predict independent tasks of personal relevance (Experiment 1) and value-based decision making (Experiment 2) will elicit shared neural substrates in the pgACC and vS as demonstrated by Enzi et al. (2009), and that this activity will be supplemented by mutual activity in the mPFC more broadly. The anticipated results constitute grounds for an inference on the mutually implicated neural correlates of self-relevance and value-based decision components that will subsequently be interrogated in Experiment 3.

In Experiment 3, if participants behave like rational economic agents, and choose the option associated with more gold coins whenever possible, then, because the total value choices is balanced across conditions, any differences in gold coins earned should indicate a subjective preference for the disclosure condition associated with more gold coins (owing to more frequent selection when the value of each option was equivalent). If we control for the monetary outcome associated with each trial, then the neural activity in functionally localized regions of interest associated with personal relevance and value-based decision making can be reasonably assumed to index the neural correlates of subjective value associated with that disclosure event.

The enhanced precision that contrasting exclusively self-referential stimulus categories against each other, coupled with a targeted search space and quantitatively differentiable indices of fiscal value provide relatively robust support for subsequent inferences about the representation of personal relevance in terms of value assignment.

It should be noted that while whole brain analyses will be conducted as a matter of course, the most relevant tests are those in the functionally defined regions of interest, which will likewise be interrogated in a similar manner. However, assuming that the anticipated regions of interest are not evident in conjunction analysis, additional results will be interrogated from anatomically defined regions appropriate for the vS, vmPFC, and pgACC. If there is a main effect of prospective social context on vS or CMS BOLD signal, then this is evidence that prospective social context modulates the neural representation of self. However, if there is no omnibus difference, I predict that comparing sharing a fact, collapsed across parent and friend, should be associated with stronger BOLD signal than keeping a fact private. I specifically predict that disclosing to friends should be associated with the strongest BOLD signal, followed by disclosing to parents, with the weakest vS and vmPFC bold associated with keeping facts private. Barring a linear pattern, I predict that prospective disclosure contexts will, collectively, elicit stronger vS and vmPFC bold signal compared to private disclosures, This should serve as a neural index of the *value* associated with *prospective* disclosure, extending findings concerning the immediate reward associated with disclosures in the present. If vS and vmPFC BOLD signal can be reliably dissociated across disclosures to parents and to friends, then this should afford the inference that the social selves elicited by the prospective disclosure context are differently valued, which should represent an entirely novel contribution.

CHAPTER III
METHODS

Participants

Participants were recruited via fliers posted around campus and screened for eligibility in a manner approved by the Institutional Review Board at the University of Oregon. Two subjects exhibited neurological abnormalities profound enough to be evident in the raw functional images at the time of acquisition, and were thus excluded from further analysis, resulting in a sample consisting of 25 right-handed, first-year college students (11 women; age M = 19.10, SD = 0.36 years) for whom neuroimaging and behavioral data were collected and analyzed. Prior to scanning, participants were informed that one of the neuroimaging experiments involved answering questions about oneself in private or social contexts, and were asked to choose a gender-matched friend that they had met since coming to college as well as a parent of their preference with whom they felt comfortable sharing trivial, yet personal information about the self via email. The differential self disclosure (DSD), probabilistic decision making (PDM), and self versus change (SVC) tasks were described to subjects as the “sharing task,” “alien identification task,” and “self or change task,” respectively. Participants practiced all three tasks in a scanner simulator until they were capable of providing timely, on task responses while refraining from making large or sudden movements (as assessed by experimenter on the basis of visual inspection).

Procedures

Experiment 1: Self versus change paradigm

In the self versus change task, participants were shown a series of 48 trait adjectives, each presented for 3.5s and accompanied by an icon indicating instruction condition (i.e., “Describes me?” or “Can it change?”) and icons prompting “yes” or “no” button presses (**Figure 4**). Twenty-four blocks consisting

of 4 trials each were presented, each preceded by a 3.5s presentation of the relevant icon and text reminding subjects of the associated instruction. Presentation of each trait adjective was jittered by 0.47s - 1.74s, during which the condition-relevant icon remained on screen. The start of each new block was preceded by a blank screen presented for 3.97s - 23.88s. An optimized presentation sequence was determined through use of a genetic algorithm designed to obtain maximal contrast detection between the two conditions of interest, as well as between adjectives describing “prosocial” or “antisocial” popularity traits, not further discussed here (Kao, Mandal, Lazar, and Stufken, 2009). Each of 48 trait adjectives was presented once per condition (96 total trials). Participants completed two functional runs, each of which lasted 6m 18s.

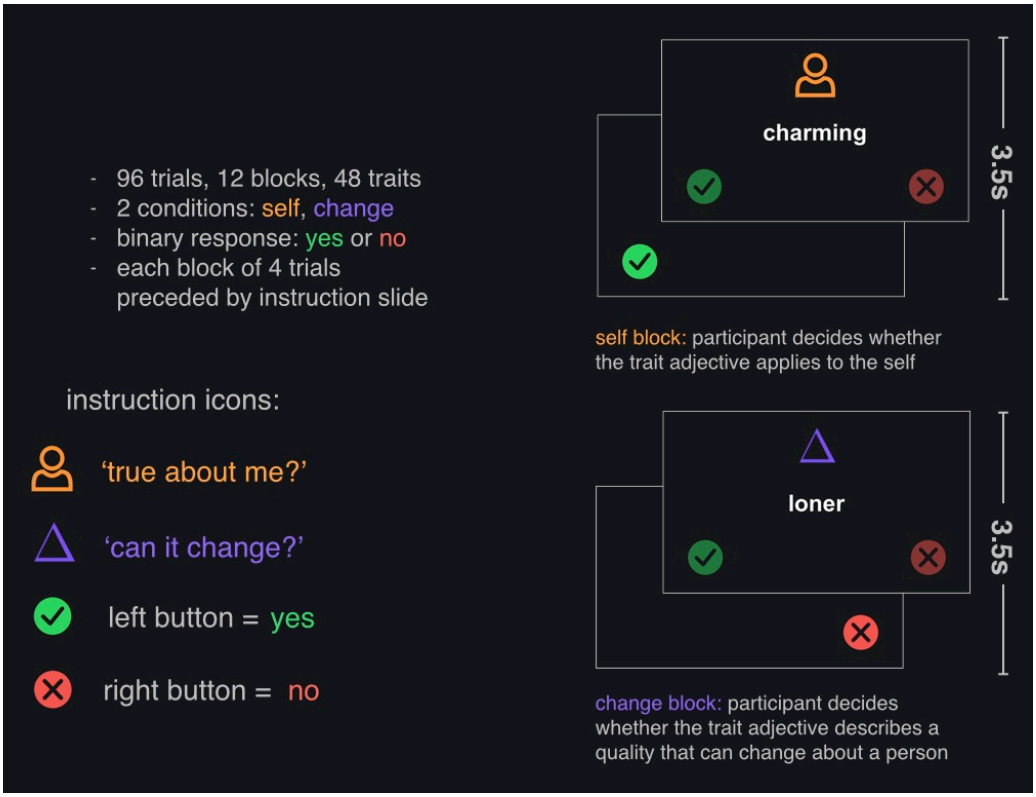

Figure 4. Self versus change (SVC) paradigm used to assess self-referrential processes in Experiment 1. Participants are presented with a trait adjective and an icon that instructs them to evaluate the trait adjective in terms of the self (i.e., ‘true about me?’) or malleability (i.e., can it change?’). Subsequent yes or no responses are collected via left or right button press, respectively.

Experiment 2: Probabilistic decision making paradigm

In the “alien identification task” (adapted from Cohen et al., 2010), participants performed a probabilistic decision making while undergoing BOLD fMRI. On each 5s trial (range 3.9s - 7.75s), participants saw an abstract, computer generated fractal stimulus that they were told represented an alien organism (**Figure 5**). They were asked to classify the alien as a member of the extra-terrestrial species “Lux” or “Raz”. Participants were told that the task was probabilistic in nature, and additionally instructed that because our ability to visually identify these aliens is not perfect, sometimes an alien that looks like one species actually belongs to the other, and therefore to expect that feedback may not be consistent for each stimulus. Stimuli were presented for an average of 3s (jittered between 2.5s and 5s), during which time participants had to indicate whether the alien was a Lux or a Raz via left or right handed button press, respectively. After participant response, feedback was presented for 1.25s, consisting of the intended response (i.e., which species the alien belongs to) as well as a reward of gold coins. The reward was either 2 or 4 gold coins, and after the task participants were paid 1.647 cents for every gold coin amassed across experiments. The intertrial interval was 0.75s on average (range 0.15s - 1.5s).

The trial order and length were optimized for separating the neural response to stimuli from the neural response to feedback, and 144 trials were spread over two 306 second runs. There were six distinct alien stimuli, two of which were predictably Lux (83%), two predictably Raz (83%), and two of which were random (50%). Across each of the three stimulus types (predictable Lux, predictable Raz, and random), one stimulus was associated with a large reward of four coins and one with a small reward of two coins. Participants completed two functional runs, each of which lasted 6m 3s.

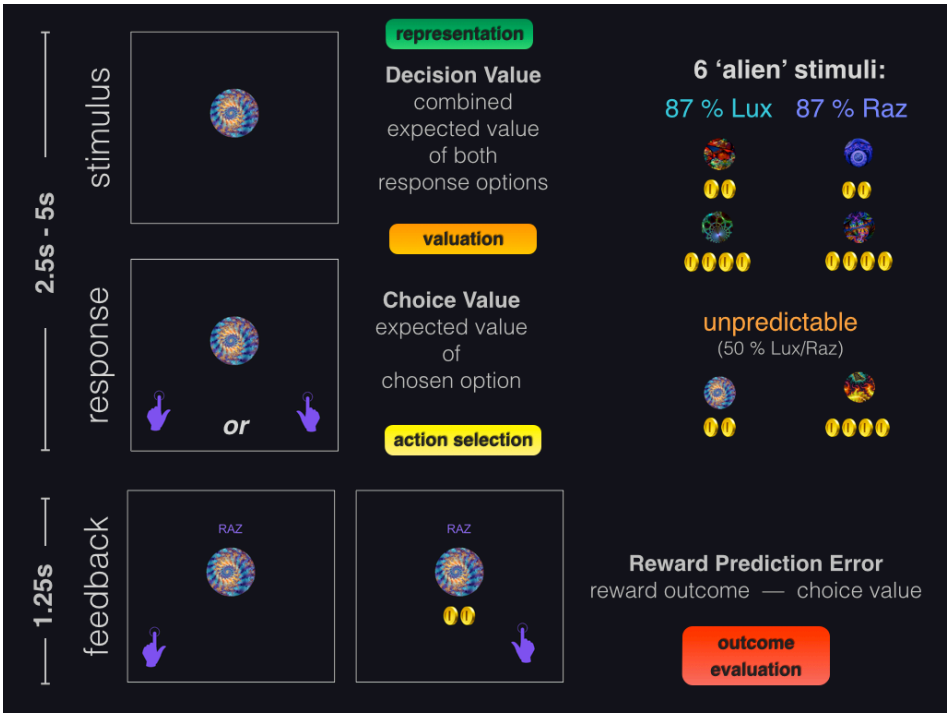

Figure 5. Probabilistic decision making paradigm used to assess reward-related processes in Experiment 2. Participants are presented with a stimulus on each trial, guess as to its 'alien identity' via left or right button press, and then receive feedback about the correctness of their response and the associated reward. The decision value (DV), choice value (CV), and reward prediction error (RPE) components are described at each phase of the experiment as they relate to the associated biological computations. It should be noted that computations associated with learning impact all other depicted stages of value-based decision making.

Experiment 3: Differential self-disclosure paradigm

In the differential self disclosure task, participants made binary choices between pairs of three possible disclosure audiences: the self ("keep it private"), a friend ("share with friend") or a parent ("share with parent"), with the social targets selected by the participant upon enrolling in the study. Each choice was associated with zero to four gold coins, (paid out at a rate of 1.647 cents per coin upon completion of the experimental session). On each trial, after choosing their preferred disclosure target / gold coin option via left hand or right hand button press, participants were presented with a trivial statement about the self to which they responded "yes" or "no" via left or right hand button press, respectively

(**Figure 6**). Importantly, participants were asked to decide on the disclosure audience before seeing each fact, in order to rule out any influence of disclosure content on chosen audience. The facts participants disclosed consisted of trivial personal statements (e.g., “I want to learn to surf,” “I hate being sick,” or “I always carry chapstick”). Participants made an audience choice and a self-disclosure statement on each of 90 trials, with presentation sequence optimized to obtain maximal contrast detection between pairs of disclosure targets and numbers of gold coins associated with each choice. Trials lasted an average of 8.5s (range 8.2s - 8.8s) and were separated by presentation of a blank screen (M = 1.28s, range 0.5s -9.78s. The choice and disclosure phases of each trial were separated by an average of 0.5s, jittered about 0.3s - 0.7s. Participants completed two functional runs, each of which lasted 7m 30s.

Figure 6. Differential self-disclosure (DSD) paradigm used to explore differences in neural correlates of personal relevance across prospectively social or immediately private contexts in Experiment 3. Participants first decide to whom they will disclose the upcoming fact based on personal preference and associated gold coins. Next, participants evaluate the applicability of trivial yes or no statements about the self in a prospective social (or immediate private) disclosure context.

Neuroimaging Data Acquisition and Image Processing

All data were acquired on a 3T Siemens Skyra MRI scanner at the Robert and Beverly Lewis Center for Neuroimaging at the University of Oregon, including T1-weighted (MP-RAGE) anatomical images as well as six functional runs (two per task) of blood oxygen-level dependent, echo-planar images (BOLD-EPI) using GRAPPA parallel accession with an acceleration factor of 2 and a multi-band acceleration factor of 3. It bears noting that multi-band slice acquisition and parallel imaging techniques enable drastic increases in the temporal or spatial resolution of functional images. Because the nature of the BOLD response signal is the rate limiting step insofar as temporal resolution, we kept the acquisition time at a relatively standard 2 seconds, while effectively doubling the number of slices and increasing number of voxels per slice in order to obtain 2mm x 2mm x 2mm isotropic voxels. Although functional neuroimaging results are often artificially resampled to this resolution (which introduces the potentially confounding effects of interpolation), whole brain coverage at this resolution for acquired BOLD-EPI data is a relatively recent innovation. Additional sequence parameters: TR = 2000ms, TE = 30ms, field of view = 200mm, matrix size=100x100), 72 oblique slices, slice thickness = 2mm, flip angle = 90o. DICOM images were converted to NIfTI format via MRIConvert (*http://lcni.uoregon.edu/~jolinda/MRIConvert/*) and non-brain tissue was removed using FSL's Brain Extraction Tool (Smith, 2002). All subsequent image processing was carried out in SPM12. For each participant, all functional volumes were realigned to the first image in the series. The effects of translational and rotational motion on signal to noise ratio were calculated using the art toolkit in SPM12 and a nuisance regressor was constructed for any volume with a displacement of more than 2mm or a change in global signal intensity of more than nine standard deviations above the mean (as compared to the previous image). The anatomical image was then placed in registration with the realigned functionals, and reorientation parameters were manually derived and applied to all images so as to set the origin above and behind the anterior commissure. Anatomical images were segmented into six tissue types using the

unified segmentation approach (Ashburner and Friston, 2005). Deformations fields from this transformation were subsequently used to warp functional images into a standard space (MNI-152 ICBM template) at 2mm isotropic resolution. Finally, functional images were then smoothed with a 4mm (FWHM) smoothing kernel and concatenated into a single 4D time-series for each of the three tasks. It should be pointed out that the high resolution of warped functional volumes does not presume an undue specificity, but rather precisely reflects the resolution at which images were sampled.

Data Analysis

For all tasks, condition effects were estimated in SPM12 using a canonical hemodynamic response function, high pass filtering (128s), correction for serial autocorrelation (AR1), and a subject specific explicit mask. Masks were calculated for each subject by intersecting optimally thresholded mask of all functional images (Ridgway et al., 2009) with the grey matter tissue probability map from unified segmentation, binarized to exclude all voxels with a less than 1% probability of being grey matter. It should be noted that while these masks exclude a small number of in-brain voxels (primarily in the ventricles and large white matter tracts), they are not stringent “grey matter masks” in the classical sense, and they contain more than 150,000 voxels on average. These individual subject masks were averaged (and re-binarized) to create an explicit mask for use in group level analyses. In order to appropriately threshold statistical results, images were either subjected to whole-brain FWE correction in SPM12 with the default extent threshold of $k \geq 5$ voxels, or else the appropriate cluster defining threshold to obtain an FWE-corrected alpha level of $p < 0.05$ was determined for each SPM using corrclusth.m (Nichols 2015, http://www2.warwick.ac.uk/fac/sci/-statistics/staff/academic-research/nichols/scripts/spm).

Model specification, Experiment 1 (self versus change)

Condition effects for self (“Describes me?”) and change (“Can it change?”) were estimated according to the general linear model in SPM12. Reaction time

(RT) on each trial was entered into the model as the duration of the event in order to control for potential discrepancies in BOLD signal due to differences in RT both within and across conditions. Instruction events were also convolved with the HRF and modeled as regressors of no interest, in addition to any trials for which participants failed to respond. Nuisance regressors were appended to the design matrix in order to partial out variability in BOLD signal due to participant motion. To assess population level effects, voxelwise statistical parametric maps summarizing the contrast between self and change trials were calculated for each participant and then entered into a random effects (group level) one sample t-test.

Model specification, Experiment 2 (probabilistic decision making)

All behavioral data were analyzed in MATLAB-R2015a (according to the method described in Cohen et al., 2010). For each subject, decision components were computed using a fully connected neural network model with one input node per stimulus and two output nodes. Connection weights were updated after each trial according to the Rescorla-Wagner rule. A learning rate (alpha), which ranged from 0 to 1, was the sole free parameter in this model, and described the extent to which individual trials modulated the association between stimuli and responses. The decision value (DV) of the presented stimulus was updated after each trial by function of the participant's learning rate (alpha) and prediction error (PE) on that trial:

$$DV(n+1) = DV(n) + (alpha) + PE(n)$$

Choice value (CV), the value associated with the participant's chosen stimulus category (i.e., Lux or Raz) was updated on each trial with the above equation, while decision value (DV) was calculated as the sum of the decision values relating to both possible outcomes for a stimulus, and was not updated on a per trial basis. Reward prediction error (PE) was computed as the difference between the actual observed outcome and the choice value on each trial. Condition effects for stimulus, response, and feedback were estimated according

to the general linear model using SPM12's canonical hemodynamic response function, high pass filtering (128s), correction for serial autocorrelation (AR1), and an optimally thresholded explicit mask (Ridgway et al., 2009). Importantly, an additional parametric modulator was convolved with the hrf for each event type: DV at stimulus onset, CV at response, and PE at feedback. The duration of stimulus events (as modulated by DV) was set to the amount of time between stimulus presentation and participant response (i.e., reaction time), duration of all response events was set to zero, and duration of all feedback events was 1.25s. Bivariate correlations between reaction time and coins earned as well as reaction time and prediction error for each trial were assessed for each subject, and no significant correlations survived correction for multiple comparisons. Nuisance regressors were appended to the design matrix in order to partial out variability in BOLD signal due to participant motion. To assess population level effects, voxelwise statistical parametric maps summarizing the effects of the parametric modulators (variability explained by CV, DV, and PE) were entered into in 1 x 3 repeated measures ANOVA, assuming non-independence and unequal variance across conditions.

Conjunction across Experiments 1 and 2

In order to formally assess overlap in the neural substrates associated with independent tasks of self- and reward-related processing, the linear contrast images for prediction error and self > change from each subject's first-level models were entered into in 1 x 2 flexible factorial model, with one regressor for PE, one for self > change, and an additional regressor no interest for each subject, assuming non-independence and unequal variance across conditions. Group level contrasts for self > change and PE error were then tested against the conjunction null hypothesis that either or both the effects of self or reward are null at each voxel, applying a height threshold of $p < 0.005$ and an extent threshold of $p < 0.05$ ($k \geq 135$ voxels, FWE corrected at the cluster level). It should be noted that specification of this model (which is essentially a paired-samples t-test) is not strictly necessary in order to test the conjunction, but for the sake of consistency

in the application of thresholds, correction for non-independence within subjects, and reporting of results, linear contrast images were organized in this fashion. Thresholded clusters from the conjunction were saved as binarized masks for use as functionally defined regions of interest (ROIs) for subsequent analyses.

Model specification, Experiment 3 (differential self-disclosure)

Neural activity associated with self disclosure was estimated on a per trial basis at the individual subject level for each of the three audiences (self, parent, and friend). In order to control for variability attributable to discrepancies in the gold coins either available or earned on each trial, each event type was parametrically modulated by the number of gold coins received. Choice events were not specified in the model, in order to partially control for activity associated with perceptual or motor processes common to choice and disclosure events, as well as to strengthen subsequent inferences based on simple effects of disclosures for each audience condition by using an “active implicit baseline” (which is not strictly ideal, but is assuredly preferable to contrasting the neural activity associated with any meaningful psychological process against staring at a fixation cross). Parameter estimates for each condition were extracted from the functionally defined ROIs via the SPM12 Volumes utility. Summary statistics was created for each subject by interpolating parameter estimates over all voxels in each mask with a 3rd degree b-spline.

Neuroinformatics Approach

Background

The nature of value computation is an active area of research in psychology, neuroscience, economics, and computer science, but as traditional disciplinary boundaries between these once disparate fields continue to erode, experiments that simultaneously address multiple levels of analysis are on the rise. Consequently, comparing and synthesizing findings from the massive datasets that ensue poses a major challenge. Distributed efforts to facilitate the organization, analysis, and sharing of these data have given rise to the

burgeoning movement known as neuroinformatics (Yarkoni, Poldrack, Van Essen, and Wager, 2010). Applying techniques from neuroeconomics to experimental design and data analysis coupled with the use of neuroinformatics tools to compare and contrast findings with those in the empirical record represents a powerful approach to supplement traditional, contrast-detection-based fMRI with formal reverse inference, rather than the brand of invalid, post-hoc conjecture that has, unfortunately, saddled "reverse inference" with the connotation of an aspersion to be avoided at all costs (Poldrack, 2011)

The typical neuroimaging experiment involves manipulating a psychological variable of interest and localizing the effects of that manipulation on neural activity, a strategy referred to as forward inference (Henson, 2006). Unfortunately, the tendency to observe neural activity in a structure and reason backwards, assuming that the psychological constructs being tested must involve cognitive processes frequently associated with activity in that structure, is a pernicious practice known as "reverse inference error" (Poldrack, 2006).

Formalizing reverse inference with Neurosynth and Neurovault

Reverse inference errors are by no means limited to neuroimaging investigations, but fMRI researchers are particularly vigilant against them, especially with respect to anatomical regions that are implicated across a wide range of cognitive tasks (e.g., mPFC, insula, amygdala). This situation is unfortunate, because while reverse inference error (or any form of post-hoc conjecture) has a negative impact on science insofar as promoting unsound conclusions and neuroimaging practices, the reverse inference baby is often cast out with the erroneous bathwater. Fortunately, NeuroSynth, a platform for large-scale data analyses and comparison of results for empirical investigations against the reported neuroimaging literature studies, allows investigators to bolster conclusions and generate new hypotheses with a formal mechanism for reverse inference (Yarkoni, Poldrack, Nichols, Van Essen, and Wager, 2011). This freely-distributed, open-source tool uses automated, text-based analysis of

over 10,000 fMRI studies to generate reverse inference maps of conceptually relevant terms by cross-indexing frequently occurring words in neuroimaging manuscripts with reported MNI-coordinates. While a forward inference map is essentially a demonstration of the probability of activity in a voxel given the presence of a conceptual term of interest, a reverse inference map indexes the probability of a term being associated with the brain-wide pattern of neural activity.

Although the peak-based meta-analyses that can be carried out in NeuroSynth are not as informative as image-based meta-analytic approaches, obtaining unthresholded statistical results for thousands of experiments is computationally impractical, as well as being a major pragmatic challenge, and this database features over 3000 relevant terms of interest (appearing in at least 20 articles) distributed over 10,000 empirical reports. Another neuroinformatics tool that grew out of NeuroSynth is called NeuroVault: a public repository of unthresholded brain activation maps (Gorgolewski et al., 2015). NeuroVault aims to overcome the pitfalls of coordinate based meta-analysis by allowing researchers to upload their unthresholded SPMs (or even the entire group-level analysis) for the purposes of easing long-distance data-sharing and -visualization, as well as facilitating distributed, collaborative, open-source neuroimaging investigations. Using NeuroValut and NeuroSynth in tandem, researchers can upload their unthresholded statistical results and “decode” them against the NeuroSynth database, generating a list of the most closely related term-based reverse inference maps. While it is important that these tools be used with care, as it is still entirely possible to use them to make an erroneous reverse inference, the use of decoding approaches to supplement (rather than replace) results based on whole-brain analyses and a priori hypotheses about neural structures and psychological concepts of interest is a powerful tool for synthesizing results from empirical investigations with results reported in the literature. Neuroinformatics perspectives and tools enhance our collective ability to compare and contrast psychological processes and neural activity that may be elicited across different kinds of experimental paradigms while controlling for the

base rate at which psychological constructs or anatomical structures are typically reported in the sorts of tasks common to neuroimaging experiments.

CHAPTER IV
RESULTS

Neuroimaging Results, Experiment 1 (Self Versus Change)

Conducting a one sample t-test at the group level over individual subject contrast images for self > change reveals robust activity in both anterior and posterior cortical midline structures, as expected (**Figure 7**, **Table 1**). A solitary supra-threshold cluster dominates the entire rostro-medial aspect of the PFC, encompassing more than 4000 voxels and extending from the the posterior ventral striatum to the frontal pole, including large portions of the paracingulate gyrus, perigenual, subgenual, and subcallosal anterior cingulate cortices (pACC, sgACC, scACC). To attain a higher degree of anatomical specificity, application of a more stringent threshold (brain-wide FWE) elucidates two distinct clusters in the mPFC: an 87 voxel cluster immediately anterior to the genu of the corpus callosum (pgACC; MNI peak mm [-6 36 4]), and a 98 voxel cluster in the anterior aspect of the paracingulate gyrus (MNI peak mm [-6 52 12]), anterior and slightly superior to the first. The cluster of activity in pgACC is quite similar to that described for the interaction between ipseity and and self-similarity in Moore et al., 2014, and this region specifically has been routinely implicated in self-specific social cognition (Northoff et al., 2006). Recruitment of cortical midline structures (CMS) for self-referential processing is by no means a finding without theoretical or empirical precedent, as scores of empirical investigations and several large-scale meta-analyses have suggested that CMS are essential to the brain's representation of self (Northoff et al., 2006; Van Der Meer, Costafreda, Aleman, and David, 2010b; Van Overwalle, 2009). However, the profound magnitude and extent of cortical midline BOLD activity in these data not only underscores the robust manner in which the current results replicate consistent findings in the social neuroscience literature, but also provides a powerful analytical tool for

further decomposition of self-relevant cognition in the independent task of self-disclosure.

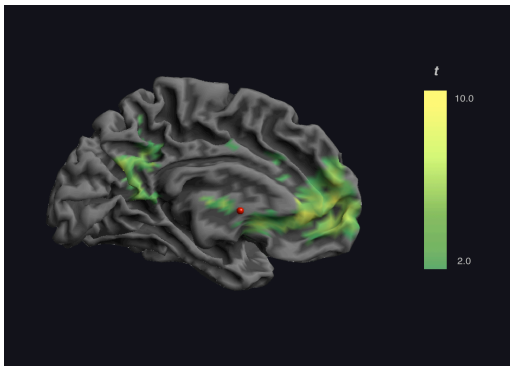

Figure 7. Group level (*N* =25) whole-brain SPM for one sample t-test of self versus change contrast at individual subject level. Red sphere indicates origin at MNI coordinates [x = 0, y = 0, z = 0] mm, thresholded for display at $p < 0.05$, (FWE corrected for multiple comparisons, extent: k ≥ 138 voxels, height: t(1,24) ≥ 2.65). For precise peak and cluster statistics, refer to Table 1.

Table 1. Peak MNI statistics for whole-brain contrast of self versus change (Experiment 1). All reported clusters are significant at $p < 0.05$, corrected for multiple comparisons at via FWE height and extent thresholds, or FWE corrected extent threshold of at least 138 contiguous voxels and uncorrected height threshold of $p < 0.005$.

	voxels	peak	MNI coordinates {mm}		
Region	**k**	***T***	**x**	**y**	**z**
perigenual anterior cingulate gyrus	87	10.73	-6	10	-8
anterior medial prefrontal cortex	98	10.28	-6	-48	28
medial posterior parietal cortex	21	8.71	-8	48	-8
ventral medial prefrontal cortex	18	8.53	-48	-68	-16
lateral orbitofronal cortex	10	7.59	24	-54	-20
posterior cingulate gyrus	5	7.35	8	-86	-10
Extent & height thresholds - $p < 0.05$ FWE					
perigenual anterior cingulate gyrus	4209	10.73	-6	36	-4
medial posterior parietal cortex	1699	8.71	-6	-64	24
lateral orbitofronal cortex	315	7.59	30	14	-20
precentral gyrus	932	6.80	36	-22	54
cerebellum	374	6.51	16	-46	-20
temporal fusiform gyrus	1065	6.45	40	-58	-16
amygdala	194	5.56	30	-4	-24
lateral occipital cortex	605	5.29	32	-86	0
occipital pole	166	5.15	4	-16	38
lateral occipital cortex	228	4.85	42	-66	12

Extent threshold - $p <$ 0.05, FWE (k ≥ 138), ***height threshold*** - $p < 0.05$ FDR

Neuroimaging Results, Experiment 2 (Probabilistic Decision Making)

Participants earned an average of $4.12 during both runs of the RPE task (gold coins: (M = 250.48, SD = 31.27), receiving more coins for predictable stimuli (M = 2.00, SD = 0.34) than random ones (M = 1.34, SD = 0.21; t(24) = 6.89, p < 0.001). Reaction times also differed such that participants responded more quickly to predictable stimuli (M = 1.09s, SD =0.26) than random ones (M = 1.20s, SD = 0.26; t(24) = -3.73, p < 0.01). Voxelwise statistical parametric maps summarizing the effects of the parametric modulators (variability explained by CV, DV, and PE) for each subject were entered into a 1 x 3 repeated measures ANOVA, assuming non-independence and unequal variance across conditions. Linear t-contrasts for each reward component of interest were used to index value of the stimulus (decision value), expected value of the participant's choice (choice value) and prediction error (PE) independently characterize the brain's response to reward during probabilistic learning. At the stringent statistical thresholds FWE corrected p < 0.05 and p < 0.001 (height and extent, respectively; applied at the whole brain level), the test of prediction error (controlling for CV and DV) revealed robust clusters of activity in the left and right ventral striatum, precisely as hypothesized (**Figure 8**). No supra-threshold clusters were evident for decision value, and choice value was associated only with activity in the bilateral occipital poles. At a more lenient height threshold, albeit still corrected for multiple comparisons at an equivalent alpha level of p < 0.05 FWE (height p < 0.005 and extent ≥ 138 voxels), choice value was additionally associated with activity in the medial prefrontal cortex and precuneus, and well as motor cortex (**Table 2**) – this is unsurprising, given that CV modulates the only event in which participants engage in button pressing activity. Value-related activity in cortical midline structures is assuredly relevant to an investigation of neural similarities between self and reward, but because nearly identical clusters were evident for prediction error at this threshold (and

because the paradigm was optimized specifically to detect prediction error signal, rather than expected value) activity from the prediction error contrast is the best analog of “reward-relevant processing” at the group level.

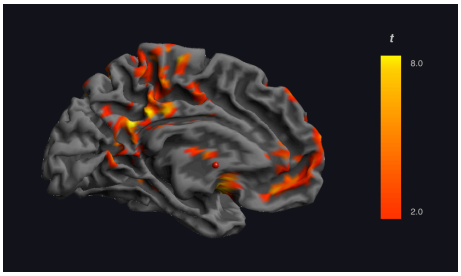

Figure 8. Group level (*N* =25) whole-brain SPM from 1 x 3 repeated measures ANOVA [*F* (2,72)] for t-contrast [*t* (1,72)] of reward prediction error, controlling for effects of decision value and choice value. Thresholded for display at $p < 0.05$, (FWE corrected, extent: k ≥ 154 voxels, height: t(1,24) ≥ 2.65).

Table 2. Peak MNI statistics for whole-brain contrast of reward prediction error (Experiment 2). All reported clusters are significant at $p < 0.05$, corrected for multiple comparisons at via FWE height and extent thresholds, or FWE corrected extent threshold of at least 138 contiguous voxels and uncorrected height threshold of $p < 0.005$.

	voxels	peak	MNI coordinates {mm}		
Region	**k**	***T***	**x**	**y**	**z**
occipital pole	512	7.9	-10	-102	10
ventral striatum	92	7.71	14	8	-10
occipital fusiform gyrus	323	7.71	38	-82	-16
occipital fusiform gyrus	249	7.23	-28	-68	-16
lateral occipital cortex	109	7.13	-26	-76	24
lateral occipital cortex	147	7.12	28	-68	30
ventral striatum	45	7.01	-14	4	-16
occipital cortex	109	6.56	48	-66	-6
inferior temporal gyrus	11	6.36	54	-56	-14
occipital cortex	16	6.15	-46	-70	-4
occipital fusiform gyrus	10	5.95	20	-82	-22
Extent & height thresholds - $p < 0.05$ FWE					
occipital pole	38692	8.09	12	-100	8
orbitofrontal cortex	321	6.07	-22	30	-18
frontal pole	391	5.27	46	44	2
cerebellum	202	5.21	32	-40	-42
superior temporal gyrus	487	5.08	-50	40	10
orbitofrontal cortex	149	4.73	24	28	-18
postcentral gyrus	1256	4.51	-58	-10	28
superior frontal gyrus	258	4.43	24	32	48

Extent threshold - $p <$ 0.05, FWE (k ≥ 138), ***height threshold*** - $p <$ 0.05 FDR

Although there is no dearth of fMRI experiments that report activity in the nucleus accumbens (a Google Scholar search for “nucleus accumbens fmri” returns > 20,000 articles as of May 10, 2015), the resolution of BOLD fMRI has, historically, been insufficient to support strong inferences about activity in nAcc proper. The nucleus accumbens is a relatively small structure and one that can be notoriously difficult to isolate from adjacent basal forebrain nuclei (e.g., caudate, putamen) even with approaches that afford a far higher degree of anatomical precision than fMRI (e.g., histological slice preparation). While high-resolution anatomical images (typically [1 x 1 x 1] mm) may allow for meaningful parcellation of the striatum into its component nuclei, classic BOLD-EPI cannot typically obtain whole-brain coverage at a resolution finer than [3.125 x 3.125 x 4] mm, and so to localize fMRI signal to the nAcc is to make a bold claim, anatomically speaking. However, the exponential increase in spatial resolution of BOLD-EPI afforded by multi-band acceleration and parallel imaging, which can acquire whole brain volumes at [2 x 2 x 2] mm in under two seconds, make inferences about function of the nAcc from BOLD-EPI data far more tenable. Granted, this increase in precision, albeit significant, still does not yield the high resolution of structural MRI, and a hasty marriage of neuroanatomical labels and blobs from statistical maps is never likely to end well, regardless of how fine-grained the images may be (Devlin and Poldrack, 2007). That said, peak activation coordinates often reported in the literature as “ventral striatum” may fall into one of several distinct structural nuclei (e.g., caudate, putamen).

In the instant case, the significant activity associated with prediction error is located almost entirely in the coordinate space of the nucleus accumbens, slightly extending into the ventral putamen in the lateral extreme and subcallosal cortex in the inferior extreme. What may appear to be a neuroanatomical digression is actually quite salient to any neuroimaging investigation of reward, as the computational models we typically use to decompose reward processes

are almost exclusively descended from models based on single unit recordings from the nAcc in slice preparation or animal models.

Conjunction across Experiments 1 and 2

A casual visual inspection of the maps for significant activity elicited by the SVC and PDM tasks implies considerable overlap between self-relevant and reward-related processing at the neural level, as we would expect in light of previous empirical demonstrations to that effect (Tamir and Mitchell, 2012). In order to formally assess this relationship, the linear contrast images for prediction error and self > change from each subject's first-level model were entered into a 1 x 2 flexible factorial model, with one regressor for PE, one for self > change, and an additional regressor of no interest for each subject, assuming non-independence and unequal variance across conditions. Group level contrasts for self > change and PE error were then tested against the conjunction null hypothesis that either or both the effects of self or reward are null at each voxel, revealing significant clusters of activity in the mesial ventral striatum, ventral medial prefrontal cortex, and both anterior and posterior aspects of the cingulate gyrus (**Figure 9**, **Table 3**). Additional clusters significant across tasks of self and reward are present in visual and motor cortex, but as these are readily explained by the perceptual and motor demands common to both tasks and not immediately relevant to the scientific questions of interest, activity in these regions is not further considered here.

Behavioral Results, Experiment 3 (Differential Self-Disclosure)

Participants earned the most gold coins for disclosures to friends (M = 83.52, SD = 18.98), followed by disclosures to parents (M = 77.84, SD = 16.44), and earned the fewest gold coins for keeping answers private (M = 63.66, SD = 31.2; $F(2,74) = 3.59$, $p < 0.05$; **Figure 10**). The average monetary value of each decision, however, did not differ across disclosure audiences ($F(2,74) = 0.045$, $p = 0.96$). This is because participants typically chose the more valuable option, but when gold coins were equal, chose parents on roughly half of available choices

(M = 50.8%, SD = 0.063), chose friends more frequently (M = 56.8%, SD = 0.15), and chose to keep answers private less often (M = 40% SD = 0.22; F(2,74) = 5.173, p < 0.01; **Figure 11**). These behavioral results were crucial in guiding the decision to include the number of gold coins received (but not the difference between gold coins received and the alternative option) as a parametric modulator in the neuroimaging models, and to variance weight the betas accordingly in light of the discrepancy of disclosure events per audience.

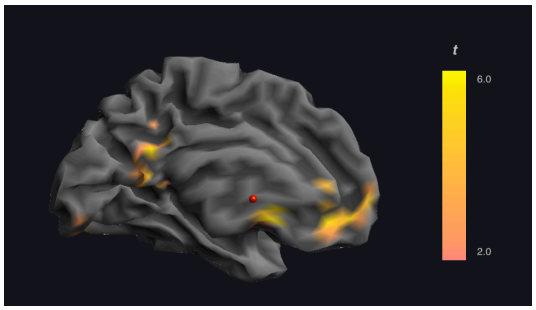

Figure 9. Group level (*N* = 25) whole-brain SPM for conjunction across contrasts from independent tasks of self-relevant cognition (Experiment 1) and reward prediction error (Experiment 2; FWE corrected for multiple comparisons, extent: k ≥ 158 voxels, height: $t(1,48) \geq 2.65$, p <0.005). Red sphere indicates origin at MNI coordinates [x = 0, y = 0, z = 0] mm thresholded for display at $p < 0.05$. For precise peak and cluster statistics, refer to Table 3.

Table 3. Peak MNI statistics for conjunction across self and reward. Coordinates for conjunction across independent tasks of self-referential cognition (self versus change contrast, Experiment 1) and probabilistic decision making (reward prediction error contrast, Experiment 2). All reported clusters are significant at $p < 0.05$, corrected for multiple comparisons via combined extent threshold of at least 138 contiguous voxels ($p < 0.05$, FWE) and height threshold of $p < 0.005$ (uncorrected).

	voxels	**peak**	**MNI coordinates {mm}**		
Region	**k**	***T***	**x**	**y**	**z**
mesial ventral striatum	347	5.63	-6	10	-8
posterior cingulate gyrus	893	5.49	-6	-48	28
perigenual anterior cingulate gyrus	1157	5.19	-8	48	-8
occipital cortex	926	5.16	-48	-68	-16
cerebellum	185	5.13	24	-54	-20
lingual gyrus	521	4.71	8	-86	-10
postcentral gyrus	317	4.47	-34	-24	52

Extent threshold - $p < 0.05$, FWE (k ≥ 158),
height threshold - $p < 0.005$ (uncorrected)

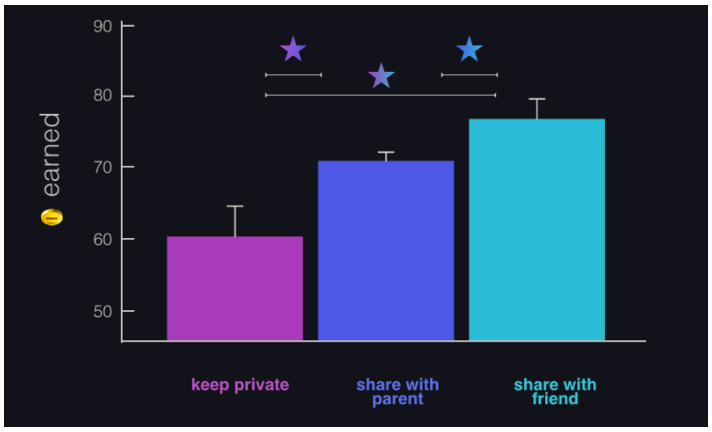

Figure 10. Coins earned by prospective disclosure condition. Participants earned significantly different amounts of coins for each disclosure audience, as indicated by a significant main effect for a 1 x 3 repeated measures ANOVA (F (2,74) = 3.59, $p < 0.05$). Stars indicate significant differences ($p < 0.05$) as indexed by post-hoc paired-samples t-tests.

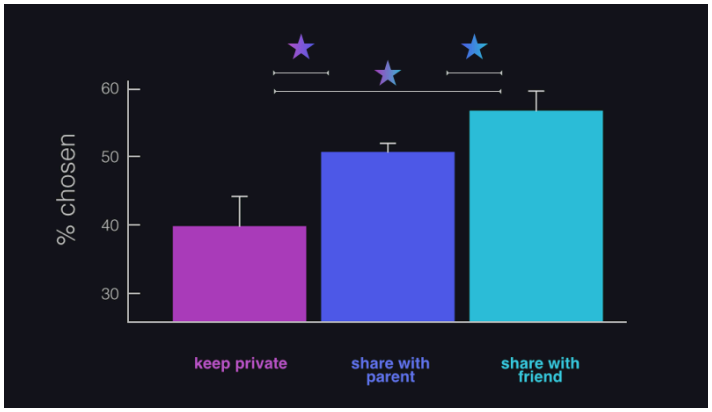

Figure 11. Effect of audience on disclosure choices. When gold coin amounts were equal, participants significantly preferred prospective social disclosures, as indicated by a significant main effect for a 1 x 3 repeated measures ANOVA (F (2,74) = 5.17, $p < 0.005$). Stars indicate significant differences ($p < 0.05$) as indexed by post-hoc paired-samples t-tests.

Neuroimaging Results, Experiment 3 (Differential Self-Disclosure)

Voxelwise statistical parametric maps summarizing the condition effects for “keep it private,” “share with friend,” and “share with parent” (controlling for monetary value of gold coins on a per trial basis) were entered into a 1 x 3 repeated measures ANOVA (with an additional regressor of no interest for each subject), assuming non-independence and unequal variance across conditions (**Table 4**). At height and extent thresholds of $p < 0.005$, an omnibus test revealed a main effect of disclosure audience in ventromedial prefrontal cortex (vmPFC) and medial posterior parietal cortex (mpPC). Subsequent contrasts (**Figure 12**) between individual audience conditions indicated that the main effect was driven by sharing (**Figure 13**).

Parameter estimates for each condition were extracted from each of the three functionally defined regions of interest (mpPC, mVs, vmPFC; clusters from the conjunction across independent tasks of self and reward) and entered into 1 x 3 repeated measures ANOVA models (**Table 5**, **Figure 14**). Responses in vmPFC and mpPC showed a linear trend, such that disclosures to friends elicited the strongest responses, followed by those to parents, and keeping answers private elicited the weakest profile of activity (mpPC: $F(2,74) = 11.56$, $p < 0.005$, vmPFC: $F = 4.55$, $p < 0.05$) . Disclosure audience also exhibited a significant effect on mesial vS activity ($F(2,74) = 5.58$, $p < 0.01$), but sharing with a parent and sharing with a friend did not differ, while keeping a fact private resulted in a (relatively) stronger deactivation.

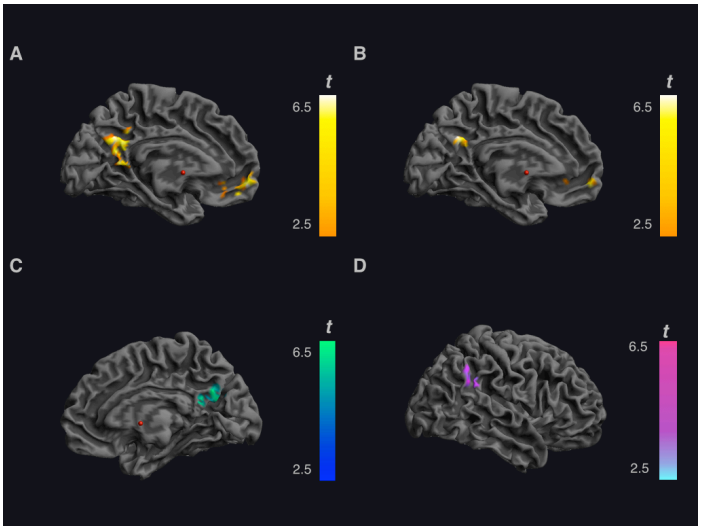

Figure 12. Whole-brain SPMs for self-disclosure contrasts. Group level (N =25) whole-brain SPMs from 1 x 3 repeated measures ANOVA [*F* (2,72)] across disclosure audience conditions (Experiment 3). Red sphere indicates origin at MNI coordinates [x = 0, y = 0, z = 0] mm. Thresholded for display at p < 0.005, (uncorrected *p* < 0.005, extent: k ≥ 75 voxels, height: *t*(1,24) ≥ 2.65). Individual t-contrasts (*t* (1,72)) across prospectively social (i.e., share with parent, share with friend) and private (i.e., 'keep it private') contexts: (A) share with friend > keep it private; (B) share with parent > keep it private; (C) share with friend > share with parent; (D) keep it private > share with parent. For precise peak and cluster statistics, refer to Table 4.

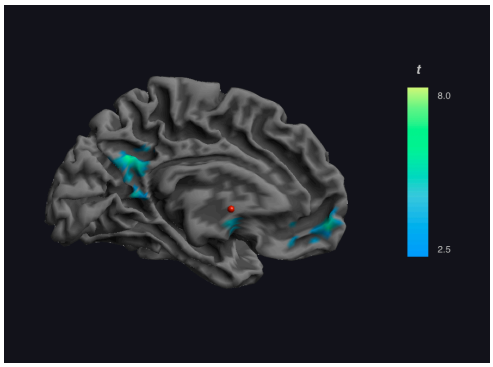

Figure 13. Whole-brain SPM for sharing versus private. Contrast of sharing (averaged across friend and parent) > 'keep it private' across disclosure audience conditions from group level (N =25) whole-brain SPM from 1 x 3 repeated measures ANOVA [*F* (2,72)] of differential self disclosure (Experiment 3). Red sphere indicates origin at MNI coordinates [x = 0, y = 0, z = 0] mm. Thresholded for display at p < 0.005, (uncorrected *p* < 0.005, extent: k ≥ 75 voxels, height: *t*(1,24) ≥ 2.65). For precise peak and cluster statistics, refer to Table 4.

Table 4. Peak MNI coordinates and stastics from group level (*N* = 25) whole-brain, 1 x 3 repeated measures ANOVA omnibus and individual condition contrast SPMs from differential self-disclosure (Experiment 3).

region	voxels	peak	MNI coordinates {mm}		
Main Effect of Disclosure Audience (F)					
	k	***T***	**x**	**y**	**z**
ventral medial prefrontal cortex	157	18.12	2	46	-18
medial posterior parietal cortex	591	15.39	-4	-56	30
share with friend > keep it private					
	k	***T***	**x**	**y**	**z**
anterior medial prefrontal cortex	281	5.29	-4	56	-10
medial posterior parietal cortex	867	5.12	0	-56	8
share with parent > keep it private					
	k	***T***	**x**	**y**	**z**
medial posterior parietal cortex	162	5.04	-4	-56	30
anterior medial prefrontal cortex	98	4.38	-4	56	-10
keep it private > share with parent					
	k	***T***	**x**	**y**	**z**
temporal parietal junction	161	5.08	54	-48	32
share with friend > share with parent					
	k	***T***	**x**	**y**	**z**
medial posterior parietal cortex	257	4.87	6	-62	26
share friend & parent > keep it private					
	k	***T***	**x**	**y**	**z**
medial posterior parietal cortex	645	5.55	-4	-56	30
anterior medial prefrontal cortex	226	5.32	4	56	-10
mesial ventral striatum	81	4.33	4	-2	-10

Extent & height thresholds - $p < 0.005$ ($k \geq 75$, uncorrected)

Table 5. Self-disclosure activity in conjunction ROIs. Differential self-disclosure activity in functionally defined regions of interest (ROIs) from conjunction across self versus change contrast (Experiment 1) and reward prediction error contrast (Experiment 2). Statistics describe repeated measures ANOVAs [*F* (2,74)] and post-hoc paired-samples t-tests [*t* (2,48)].

Conjunction across Experiments 1 & 2 {self > change && prediction error}			
mesial ventral striatum			
	F	***p***	***η2***
main effect	5.59	0.007	18.88
	T	***p***	
friend > self	3.17	0.004	
parent > self	2.39	0.025	
parent > friend	-0.39	0.697	
ventral medial prefrontal cortex			
	F	***p***	***η2***
main effect	4.55	0.016	15.95
	T	***p***	
friend > self	1.68	0.104	
parent > self	2.44	0.023	
parent > friend	1.84	0.079	
medial posterior parietal cortex			
	F	***p***	***η2***
main effect	11.56	0.001	32.51
	T	***p***	
friend > self	2.91	0.008	
parent > self	4.07	0.000	
parent > friend	2.56	0.017	

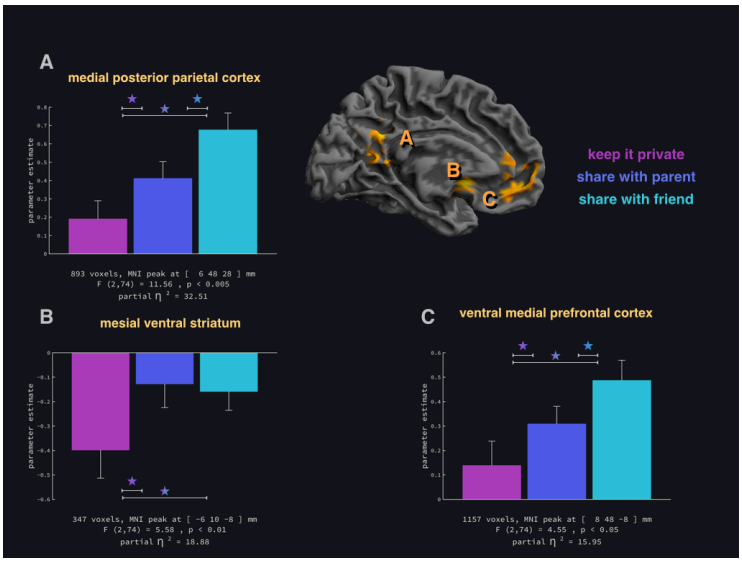

Figure 14. Self-disclosure activity in conjunction ROIs. Differential self-disclosure activity (Experiment 3) in functional regions of interest derived from conjunction across self versus change (Experiment 1) and reward prediction error (Experiment 2). Stars indicate significant differences ($p < 0.05$) as indexed by post-hoc paired-samples t-tests. For statistical results and precise MNI peak coordinates, see Tables 4 & 5.

Four additional ROIs were created from the stringently thresholded (whole-brain FWE) SPMs: pgACC and amPFC from the self > change contrast (**Table 6**, **Figure 15**) left and right vS from the prediction error contrast (**Table 7**, **Figure 16**); all results masked to exclude voxels significant in the conjunction at a height threshold of $p < 0.005$ and extent threshold of $k \geq 58$). Activity in the amPFC ROI exhibited the same significant linear pattern as other cortical midline structures ($F(2,74) = 9.20$, $p < 0.005$), and pgACC demonstrated a similar trend, but the difference hovers directly on the razor's edge of significance ($F(2,74) = 3.18$, $p = 0.050$). The left and right ventral striatal ROIs, however, do not exhibit significant differences across disclosure conditions ($F(2,74) = 1.00$, $p = 0.38$ and $F(2,74) = 0.20$, respectively).

Table 6. Self-disclosure activity in self-relevance ROIs. Differential self-disclosure activity in functionally defined regions of interest (ROIs) from conjunction across self versus change contrast (Experiment 1). Statistics describe repeated measures ANOVAs [*F* (2,74)] and post-hoc paired-samples t-tests [*t* (2,48)].

Self > change (Experiment 1)			
anterior medial prefrontal cortex			
	F	***p***	***η2***
	9.202	0.001	27.72
main effect	***T***	***p***	
friend > self	2.6131	0.015	
parent > self	3.3687	0.003	
parent > friend	2.5509	0.018	
medial posterior parietal cortex			
	F	***p***	***η2***
	3.182	0.050	11.71
main effect	***T***	***p***	
friend > self	1.4886	0.150	
parent > self	2.3314	0.029	
parent > friend	1.1312	0.269	

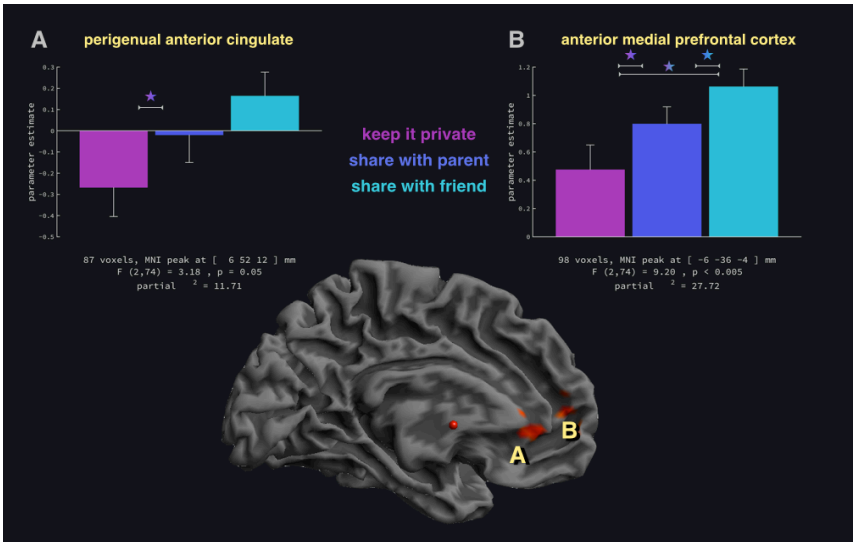

Figure 15. Self-disclosure activity in self-relevance ROIs. Differential self-disclosure activity (Experiment 3) in functional regions of interest derived from contrast of self versus change (Experiment 1). Stars indicate significant differences ($p < 0.05$) as indexed by post-hoc, paired-samples t-tests. For statistical results and precise MNI peak coordinates, see Tables 1 & 5.

Table 7. Self-disclosure activity in reward prediction error ROIs. Differential self-disclosure activity in functionally defined regions of interest (ROIs) from reward prediction error contrast (Experiment 2). Statistics describe repeated measures ANOVAs [*F* (2,74)] and post-hoc paired-samples t-tests [*t* (2,48)].

Reward prediction error (Experiment 2)			
left ventral striatum			
	F	***p***	***η2***
main effect	0.998	0.376	3.99
	T	***p***	
friend > self	1.5515	0.134	
parent > self	1.0408	0.308	
parent > friend	-0.0433	0.966	
right ventral striatum			
	F	***p***	***η2***
main effect	1.67	0.199	6.51
	T	***p***	
friend > self	1.8169	0.082	
parent > self	1.3368	0.194	
parent > friend	-0.3733	0.712	

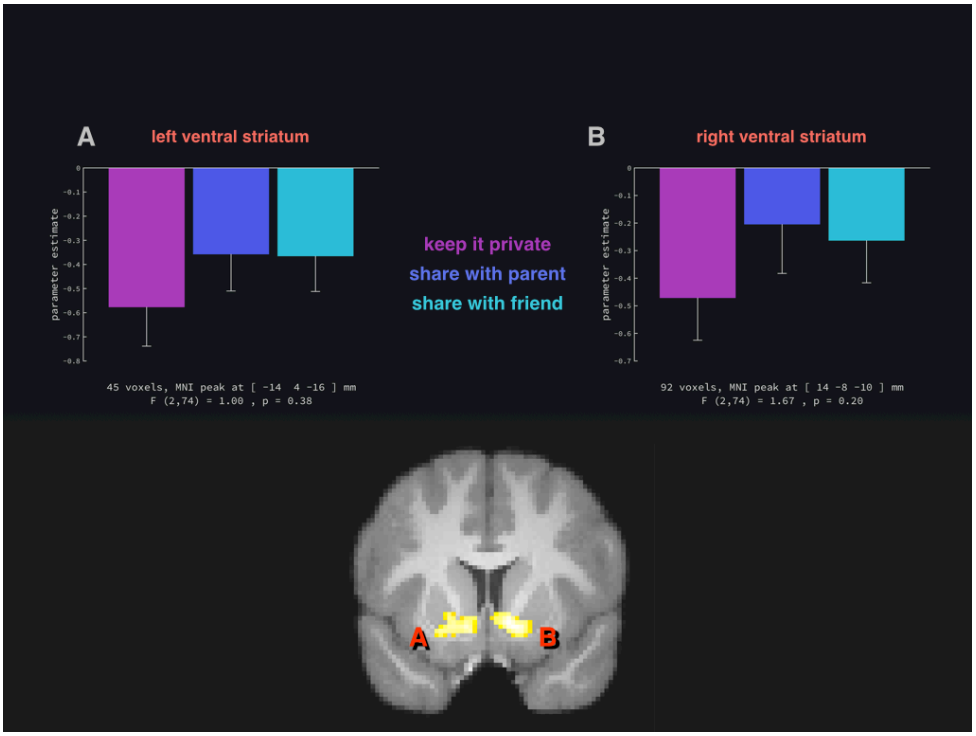

Figure 16. Self-disclosure activity in reward prediction error ROIs. Differential self-disclosure activity (Experiment 3) in functional regions of interest derived from contrast of reward prediction error, controlling for choice and decision value (Experiment 2). Stars indicate significant differences ($p < 0.05$) as indexed by post-hoc, paired-samples t-tests. For statistical results and precise MNI peak coordinates, see Tables 2 & 5.

CHAPTER V
DISCUSSION

Overview

Although the primary questions of interest in the current report center on prospective social modulation of self-relevant neural activity, Experiments 1 and 2 were specifically designed and implemented to create a conjoined search space of neural substrates shared across independent tasks of personal relevance and probabilistic decision making. Consequently, these results are briefly addressed initially in order to provide appropriate context for the discussion of region of interest based analyses of differential self disclosure (Experiment 3) activity. The impact and limitations of Experiments 1 and 2 will be considered after thorough treatment of the primary task of interest, followed by conclusions about the relevance of the current work, implications and alternative interpretations, and next steps for further advancing the state of neuroimaging investigations of the selves.

Conclusions

Personal relevance and reward prediction error signals

In Experiment 1, self-referential versus non-self-referential cognition elicited robust activity in medial prefrontal cortex (mPFC) and medial posterior parietal cortex (mPPC). This supports the well-replicated hypothesis that cortical midline activity indexes personal relevance. Support for the hypothesis that vS activity would be likewise associated with personal relevance is not clearly evident at stringent, whole-brain height and extent thresholds. Somewhat less direct support for this hypothesis is provided by evidence of significant mvS activity in the conjunction analysis, which is also visually apparent in the contrast of self versus change. Due to the massive spatial extent of significant clusters,

however, this activity cannot be attributed solely to vS (Woo, Krishnan, and Wager, 2014). In Experiment 2, I simulated probabilistic decision making in an artificial neural network for each participant, and computed a parameter for reward prediction error at each trial. Reward prediction error, formally defined by the difference between reward outcome and expected value of a probabilistic decision, accounted for robust BOLD signal in the bilateral ventral striatum (vS), in keeping with the hypothesis that the brain's reward prediction error signal is computed by midbrain dopaminergic innervation of the nucleus accumbens (Delgado et al., 2002). Although incremental replication may not yield the same reward as that of unprecedented scientific discovery, there is still merit in assessing and updating our paradigms even for processes with considerable empirical and meta-analytic precedent to guide predictions. Both the self versus change contrast from Experiment 1 and the computationally derived reward prediction error contrast from Experiment 2 elicited the hypothesized patterns of neural activity, and should be regarded as having successfully functionally localized self-relevant and RPE related activity, respectively.

Independent and overlapping neural correlates of self-evaluation and value-based decision making

Conjointly, the hypothesized functional overlap between personal relevance and prediction error signal was evident in mesial vS and vmPFC, as well as in the posterior cingulate cortex (PCC). I initially predicted that BOLD signal in pgACC would be likewise mutually elicited across personal relevance and reward prediction error in light of the pgACC reported by Enzi et al. (2009). On further reflection, however, its absence from the conjunction analysis in the current report is unsurprising, as Enzi and colleagues additionally demonstrated that the difference in pgACC activity between high personal relevance and a control task consisting of simple figural orientation judgments was dramatically higher than differences between pgACC responses to positive reward outcomes and the same, low-level control task.

Effects of disclosure audience

It should first be pointed out that no explicit, economic model was applied to describe the behavioral choices in Experiment 3. Because participants behaved like rational agents with minimal deviation from optimal behavior, the raw gold coin amounts represent the monetary value associated with each choice. This was accounted for at the neural level by inclusion of a parametric modulator that controls for the financial value of these choices.

Accounting for monetary value as described at the single subject level, group-level analysis revealed a brain-wide main effect of disclosure audience in the ventral medial prefrontal cortex (vmPFC) and the medial posterior parietal cortex (mPCC), providing fairly strong evidence that prospective social context modulates self-evaluations in cortical midline structures (CMS). Furthermore, the enhanced BOLD signal for prospective disclosures to friends compared to parents in mPPC supports the hypothesis that the neural mechanisms underlying assignment of personal relevance differentiate the prospective social contexts of future disclosures to parents as compared to friends. The operationalization of personal relevance in this paradigm is also noteworthy and strengthens the above inferences. As many neuroimaging investigations of the self that report personally relevant activity define self-relevance in terms of self-referential cognition, as contrasted against non-self-referential stimuli, which may confound the detection of effects specific to isolating the process of interest. The differential self disclosure task (Experiment 3) is unique in that it controls for self-referential processes to the extent that all disclosure statements involve explicit awareness of the objective self as a stimulus. Consequently, any differences in neural activity associated with prospective disclosure conditions may reliably be attributed to the assignment or representation of personal relevance, rather than differences in the degree to which conditions elicit self-referential processes. It is

plausible that the self-referential decision associated with each prospective disclosure might involve the computations of outcome evaluation and learning from the choice of disclosure audience, as well as representation, value assignment, and action selection during the subsequent disclosure. However, because the disclosure statements are all essentially binary personal relevance judgments, and personal relevance judgments depend on value computation, it is most likely that the self-referential cognition elicited in Experiment 3 is tied to the neural computation of value assignment or neural representation of that value.

Comparing prospective disclosures (collapsed across friend and parent) against self-referential cognition in a private context revealed activity in the mvS in addition to vmPFC, consistent with Tamir and Mitchell's (2012) finding that self disclosure is associated with stronger whole-brain vS and vmPFC responses than keeping answers private. The current work also extends this line of research by demonstrating that the effect applies not only to the reward outcomes associated with immediate social sharing about the self, but to the value of disclosures to be resolved in the future. The activity associated with prospective social selves is quite compelling in its immediate visual similarity to the conjunction across independent tasks of self-reference (Experiment 1) and reward prediction error (Experiment 2). Independent contrasts of prospective social disclosures versus private self-reference revealed patterns of activity in CMS similar to those evident in the main effect contrast, but was dramatically more robust in mPPC for sharing with friend, while clusters of mPFC activity were similar for disclosures to either friends or parents compared to private self-evaluations, but situated approximately 1cm anterior and superior to the peak of the omnibus test of social context. Although this may seem like a minute difference, the distinction between the vmPFC cluster associated with combined disclosures and the slightly more superior and anterior clusters associated with disclosures to parents and friends, individually track roughly along Northoff and colleague's (2006) proposed boundary in the mPFC between paralimbic (e.g., pgACC, vmPFC) and association cortices (e.g., amPFC, mPPC). According to this scheme, these anatomically distinct cortical regions comprise the integrative

aspect of the self, which combines visceral, homeostatic information about the internal self projected to paralimbic regions (e.g., pgACC, vmPFC) with information about the external self from primary sensory and motor projections to heteromodal association cortices (e.g., amPFC, mPPC; Northoff, Feinberg, and Qin, 2010). The individual clusters associated with friend versus private and parent versus private contexts more closely resemble the region of amPFC Nicolle et al. (2011) implicated in abstract models, while the vmPFC cluster for disclosing compared to keeping self-evaluations private more closely resembles paralimbic regions implicated in integrating interoceptive signals with externally derived sensorimotor information meaning (Roy, Shohamy, and Wager, 2013). However, while these observations are well in line with the conceptual approach of this the empirical study at hand, they are not informed by a formal *a priori* hypotheses, and are discussed to provide context and conceptual resonance with the work on which this research is based, rather than to serve as evidence for any inferences about the proposed broader function of mPFC. Collectively, findings at the whole brain level support the hypothesis that activity in the vS and CMS differentiate private self-referential cognition from the prospective social context of future disclosure (controlling for any discrepancies in reward outcomes across conditions).

Differential assignment of personal relevance and value in functionally defined regions of interest

Differential responses in the mesial ventral striatum (mvS) distinguished between both social contexts as individually compared against private self-reflection, which is consistent with Tamir and Mitchell's (2012) finding, but did not distinguish between disclosures to parents and disclosures to friends. In mPPC and vmPFC, structures mutually implicated in independent tasks of self and reward, BOLD signal increased linearly for keeping a fact private, sharing with a parent, and sharing with a friend, respectively. In concert with the visually evident similarities between the conjunction and contrast of disclosures versus private self-evaluations and results at the whole brain level, this is relatively strong

evidence that the value assignment associated with personal relevance is contingent upon the imagined social context (or lack thereof) in which stimuli are presented. Not only does it appear that the prospective social selves collectively "rank more highly" (p.74) than the isolated scanner-self, as indexed by vS differentiation between sharing and private self-evaluations, but cortical midline structures further differentiate between the personal relevance associated with prospective social contexts, supporting the hypothesis that cortical midline structures implement multiple aspects of the social self or selves. It may be the case that thinking about potential social selves elicits more robust "self-referential" activity than thinking about the isolated, but immediate self.

Implications and Next Steps

NeuroSynth, a tool for conducting formal reverse inference, allows users to upload unthresholded statistical parametric maps and "decode" them against a meta-analytic database of reverse-inference maps, automatically generated by cross-indexing frequently occurring words in neuroimaging manuscripts with reported MNI-coordinates (Yarkoni, Poldrack, Nichols, Van Essen, and Wager, 2011). Analysis of the statistical parametric map (SPM) for the group level contrast of share with friend > answer privately revealed that it was more tightly linked to reverse inference maps associated with "autobiographical," "self," "self-referential" than to any other psychological processes. The only terms more strongly correlated describe the spatial or anatomical regions active in the contrast (e.g., "medial prefrontal," "posterior cingulate," "midline") or brain-wide networks of which they are components (e.g., "default mode"). Although the aforementioned terms, conveniently, describe the regions of interest derived from the overlap of independent tasks of self and reward, the cortical midline is likewise engaged by many other psychological processes. However, because these correlations are with the reverse, rather than forward inference maps associated with the terms, they suggest that the share with friend versus answer privately contrast is *selectively* associated with those regions, controlling for their prevalence in the neuroimaging literature. A similar, albeit weaker, pattern was

evident for the group level T-map for share with parent versus answer privately. The inverse contrast (private > parent) was related predominantly to terms concerning attention, working memory, or executive control (no socially relevant terms out of the strongest 10% of correlations).

Synthesizing whole-brain findings on the effects of disclosure audience, differential activity in independently derived regions of interest, and meta-analytic decoding of the unthresholded SPMs, this work collectively suggests that thinking about the self one wants to share with a friend or parent elicits neural activity that is more similar to what is typically reported as "self-referential" processing than thinking about the self in private. Although this does not constitute strong evidence that the self is a social construct, it is an interesting finding, especially in light of William James' sentiments that we have multiple social selves, and that out of all of them, "the potential social Me is the most interesting" (p. 190), at least in terms of the allocation of neural resources.

Another means of exploring the dynamics of multiple social selves might be to distinguish true self-disclosures from evaluations of previously shared information. If guesses can be successfully retrieved from parents and friends about the yes/no answers that participants disclosed, then whether or not each item constitutes a novel "disclosure" or rather "shared self knowledge" could be determined by comparing all congruent answers (e.g., participant answered yes to "I want to learn to surf," friend guessed participant would say yes) to incongruent answers (e.g., participant answered no to "I hate spicy mustard," Mom guessed participant would answer yes). Using a state-based decision algorithm might afford the opportunity to classify neural activity associated with "disclosures" compared to "shared self knowledge" and predict behavioral choices on each trial accordingly. If a "model free" reinforcement-learning algorithm can be trained to classify "shared self knowledge" trials based on the neural signature of "friend-Me" or "parent-Me" compared to the consequence free and non-social "private-Me" (or minimally social "scanner-Me" that only involves a spatially remote experimenter and technician), this would prove a powerful demonstration of differentiable social selves.

One entirely unanticipated result is the presence of activity in the right temporal-parietal junction (rTPJ) for private self-reflection versus prospective disclosures to parents. This observation is initially confusing in light of routine implications of rTPJ in perspective taking, mentalizing, theory of mind, and other tasks of social cognition (van Overwalle, 2009), which would suggest that disclosures to friends or to parents would be more likely to elicit such a pattern, but rTPJ is also involved in a number of non-social attentional processes (Mitchell, et al., 2006), and decoding the SPM did not indicate any correlations with social terms. The reciprocally inverted patterns of CMS activity for prospective social contexts compared to private self-evaluations suggests that this anomaly may be part of a broader pattern, but one that is difficult to discern. A recent methodologically and conceptually innovative study used multivariate classification and economic models of decision behavior to classify subjects choices in a gambling game against human or computer opponents based on multivariate decoding of 110 anatomically parcellated regions (Carter, Bowling, Reeck and Huettel, 2012). Only rTPJ was uniquely capable of classifying behavioral decisions that involved both social context *and* relevant future outcomes. This illustrates a particularly compelling approach to the open question about the function of rTPJ in these processes, because it provides the chance to test whether we can classify behavioral choices based only on patterns of associated neural activity. While a model-based algorithm like the R-W model used to assess reward prediction error may be too coarse for making fine-grained social distinctions of this sort, abstract, state-based, "model free" algorithms have been used to explain vmPFC activity for tasks in which there is no "optimal" response (Hampton, et al., 2006).

Limitations and Alternative Interpretations

One factor that may confound the results reported in Experiment 3 is that, because no true "share with self in the future" condition was implemented, comparisons of prospective disclosures against private self-reflection may be contaminated by differences in effects of engaging in prospective cognition.

Although autobiographical recall and projection of oneself into the future tend to elicit similar patterns of neural activity, these patterns of activity are also located in precisely the same locations as the medial prefrontal activity observed in the present task (Andrews-Hannah, Saxe, and Yarkoni, 2014). The same differences between sharing with friend, sharing with parent, and choosing not to share could be plausibly elicited by simply engaging in the most prospective memory when thinking about a new friend (because they represent the information domain most likely to be incomplete), less when thinking about their parents (about whom considerably less prospection is likely to be elicited) and the least for private reflection, which does not require projecting oneself into the future.

Alternatively, it may not be projecting oneself into the future *per se* that drives this effect, but rather the difference in cognitive load between simply evaluating the current self and evaluating a simulated self about which to disclose. Because both prospective disclosure conditions entail future consequences while the consequences of private self-evaluation are largely resolved immediately, prospective disclosures may additional involve processes of elaboration, simulation, or imagination. Yet another interpretation is that events with implications for future outcomes are more closely attended to. However, one prominent hypothesis of attention at the cellular level describes attention in a normative framework for synaptic gain-modulation (Reynolds and Heeger, 2009) that is essentially the same as models of gain modulation that explain saccades toward a rewarding target (Louie and Glimcher, 2011). It could be likewise argued that salience is, essentially, a question of motivational or personal relevance (Schacter, et al., 2007). ‘Incentive salience’ is also one of the primary components of reward-related processes that Berridge (2012) describes as relating to motivation or desire. The point of these arguments is not to obfuscate alternative interpretations with circular logic, but rather to suggest that, because the self can be implicated in most, if not all, psychological phenomena, looking for explanations based on neurobiological mechanisms that are similarly implicated across processes of interest may prove a more fruitful approach. What this means is that if formally defined, computational models can simultaneously

explain observed behavior and characterize the neural activity underlying the BOLD signal, they may be an extremely effective strategy for addressing semantically defined psychological concepts at multiple levels of analysis (Cacioppo and Bernston, 1992; Rangel, Camerer, and Montague, 2008).

The current sample was constrained to first-year college students because we assumed that a new friend would be the most salient or motivationally relevant social context for the self during this transitional period, allowing us to better differentiate the "friend-self" from the "parent-self." It is known that early adolescents recruit stronger activity than adults in CMS during direct self-reflection, and that the extent of this activity is further modulated by the interaction of social context and stimulus content (Pfeifer, et al., 2009), conclusions based on the current results should be considered in light of the fact that a similarly enhanced response may be evident for late adolescents (i.e., the current sample), but absent in a more typically "adult" population. The results of this dissertation are consistent with other prior work in our laboratory, namely that vmPFC responses to social self-evaluations are known to increase longitudinally from late childhood to early adolescence (Pfeifer et al., 2013), and striatal responses differ for early adolescents and adults across content-based and process-based manipulations of self-referential stimuli (Jankowski et al., 2014). In light of these findings, two potential confounds should be considered. First, although all self-disclosure statements were designed to be equally trivial, the domain content of some statements could be construed as academic (e.g. "I like to read books"), while others are more obviously in the social domain (e.g., "I make people laugh"). Because statements were randomized across pairs of possible disclosure audiences, it is possible that domain-specific stimuli are more prevalent in one condition for some subjects than others. Secondly, although we assume that first-year college students will be most likely to value the social context associated with a new friend more than that associated with a parent, this may vary widely across participants. Assessment of additional self-report measures concerning the precise nature of participants' relationships to the social targets of interest may help to clarify whether the assumption that

disclosures to new friends can be regarded as more salient than those to parents.

Impact and Future Directions

A new paradigm for self-referential processing

The dominant self-reference control stimulus in Experiment 1 is the inquiry, “Can it change?” (about people in general) with regard to a social trait adjective. Although evaluating the extent to which a trait adjective is static may seem like an unlikely task to contrast against self-referential evaluations, the use of this phrase as the most appropriate control was determined via an iterative process of conceptual and empirical refinement. Empirical validation of this paradigm through traditional, forward inference approach was bolstered by a formal reverse inference, conducted by uploading the unthresholded SPM to the NeuroVault (Gorgolewski et al., 2015) repository and carrying out whole brain decoding against the NeuroSynth database (Yarkoni et al., 2011). Decoding revealed that the self > change SPM is more specifically linked to reported activity in the literature associated with the word “self” than over 3,300 other topics of interest.

By combining forward and reverse inference approaches, we can be more confident in the extent to which the operationalization of self-referential cognition in terms of the self versus change contrast reflects reports in the literature empirically, rather than by carrying out a motivated visual search for activity in structures of interest. Reverse inference must be conducted with great care, as post-hoc rationalization about unexplained activity in a neural structure in terms of the psychological processes that frequently implicate that structure assuredly constitutes a logical error. Extending support for *a priori* hypotheses by considering the probability that conceptually relevant words appear in the literature (given the empirically derived neural activity) in tandem with the probability of neural activity (given the operationalized psychological concept) is not, however, a reverse inference error, but a measured and justifiable application of formal reverse inference (Poldrack, 2011). By validating this

paradigm through traditional forward and meta-analytic reverse inference, this work provides a powerful tool for future investigations of self-referential cognition or personal relevance. In addition to the potential for giving rise to increasingly abstract and complex explorations of who people think they are or might become, these findings have practical implications as well, to be discussed in the next section.

Implications for development and psychopathology

All of the paradigms in the current work were designed with a specific eye toward ready deployment in developmental populations. Adolescents are often portrayed as making more risky decisions than adults or children (Burnett, Bault, Coricelli, and Blakemore, 2010), especially in the presence of peers (Steinberg 2008), although this has been called into question by a recent meta-analysis (Defoe et al., 2014). Social contexts also heighten adolescent preferences for immediate rewards (O'Brien, Albert, Chein, and Steinberg, 2011). Therefore, expanding disclosure audiences to include self-identified versus parent-identified "bad influences" or "good influences" may be one way to differentiate the extent to which we select our own potential future selves from the extent to which they are chosen for us. We know that the adolescent brain is more fine-grained than a phrenological seesaw in which adult-sized basal ganglia are pitted against an immature neocortex, but demonstrating that the ventral striatum contributes to ultimately wise (or at least parent-approved) decisions as well as poor ones may help better characterize this oft-maligned region of the brain as social-self seeking rather than simply wild thrill seeking (Pfeifer and Allen, 2012).

Research on self-relevance and reward may also contribute to our understanding of maladaptive behavior, and de Greck and colleagues (2008) have shown that pathological gamblers demonstrate attenuated vS and vmPFC responses to both self-relevant and rewarding stimuli. A better understanding of social influences on the brain's valuation systems will inform our academic understanding issues of substance dependence and abuse, but it may also help us to actively solve individual and societal problems caused by addiction, by

identifying at-risk populations, creating targeted interventions, or designing brain-based, personally tailored motivational strategies (Berkman, in press).

Further exploration of the default mode and self-relevance judgments may also inform our understanding of depression. Severity of depressive symptoms has been shown to correlate with perigenual cingulate responses to negatively valenced, self-relevant stimuli (Wagner et al., 2013). Although prediction-valuation models may seem to some like an overly abstract or even obtuse way of approaching psychopathology, one neuroanatomical explanation of depression and anxiety suggests that these exaggerated self-evaluative responses reflect the negatively biased updating of cognitive expectations from noisy interoceptive signals like reward prediction error (Paulus and Stein, 2010). A hypothesis that relates this anatomical framework to the self suggests that imbalanced integration between noisy, interoceptive signals, abstract affective evaluations, and external sensory information is what leads to the negatively biased self-evaluations and maladaptive expectations associated with depression (Northoff, Wiebking, Feinberg, and Panksepp, 2011). The self versus change paradigm described here may be particularly useful in identifying neural markers of depression. Although largely unexplored in the current work, the binary responses for each trait adjective can be analyzed to identify trials on which participants make negative self-evaluations for qualities that they also identify as unlikely to change. These behavioral prevalence and neural correlates of such trials may help to identify people and patterns that are at risk for depression (or perhaps even differentiate healthy and pathological selves within a single brain).

Concluding remarks

The present work replicates and extend previous findings concerning broad overlap between self and reward in the more precisely constrained contexts of personal relevance and value assignment (Enzi et al., 2009). This work also demonstrates that sharing information about the self is rewarding (Tamir and Mitchell, 2012), and that sharing about the self in a prospective context is valuable. A more precise quantification of the various people and

content that make up our future selves may lead us to a better understanding of what differentiates among the value assigned to more personally relevant, prospective disclosures to friends or parents. It may also help us to more broadly understand reward and value in terms of the immediate and long-term consequences associated with any particular aspect of the self.

A liberal interpretation of these results suggests that we differentially value the selves we are likely to become in specific social contexts. Collectively, these findings suggest that we may be able to more precisely quantify self in terms of -- well, terms. Although this may sound tongue in cheek, a comprehensive battery that assesses the trait adjectives in the self versus change paradigm with respect to as many social agents as can reasonably be elicited from subjects would provide a highly dimensional problem space that abstract-state based algorithms are well suited for, and understanding the relationship between the words we use to describe ourselves and who we essentially are may soon be not only an excellent question, but an empirical question (Alfano, 2015). Although more research is needed to extend and apply these findings, the work carried out in service of this dissertation provides important first steps as well as methodological, empirical, and theoretical contributions to the study of social influences on the self in the brain.

REFERENCES CITED

Abraham, A., Kaufmann, C., Redlich, R., Hermann, A., Stark, R., Stevens, S., & Hermann, C. (2013). Self-referential and anxiety-relevant information processing in subclinical social anxiety: an fMRI study. *Brain Imaging and Behavior*, *7*(1), 35-48.

Abraham, A., & von Kramon, D.Y. (2009). Reality = relevance? Insights from spontaneous modulations of the brain's default network when telling apart reality from fiction. *Frontiers in Human Neuroscience*, *4*(3), e4741.

Alfano, M. (2015). How one becomes what one is called: On the relation between traits and trait-terms in Nietzsche. *Journal of Nietzsche Studies*, *46*(1), 11-11.

Andrews-Hanna, J.R., Saxe, R., & Yarkoni, T. (2014). Contributions of episodic retrieval and mentalizing to autobiographical thought: evidence from functional neuroimaging, resting-state connectivity, and fMRI meta-analyses. *Neuroimage*, *91*, 324-335.

Bartra, O., McGuire, J.T., & Kable, J.W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*(1), 416-427.

Ballard, K., & Knutson, B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. *NeuroImage*, *45*(1), 143–150.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., & Rushworth, M.F.S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–9.

Berkman, E.T, in press.

Berns, G.S., Cohen, J.D., & Mintun, M.A. (1997). Brain regions responsive to novelty in the absence of awareness. *Science*, *276*, 1272-1275.

Berridge, K.C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, *35*(7), 1124–1143.

Buckner, R.L. & Carroll, D.C. (2007). Self-projection and the brain. *Trends in Cognitive Science*, *11*, 49-57.

Bush, R.R. & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313–323.

Cacioppo, J.T. & Berntson, G.G. (1992). Social psychological contributions to the decade of the brain: doctrine of multilevel analysis. *American Psychologist*, *47*(8), 1019-1028.

Carter, R.M., Bowling, D.L., Reeck, C., & Huettel, S.A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, *37*(6090), 109-111.

Coan, J.A., Schaefer, H.S., & Davidson, R.J. (2006). Lending a hand: social regulation of the neural response to threat. *Psychological Science*, *17*(12), 1032-1039.

Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., & Poldrack, R.A. (2010). A unique adolescent response to reward prediction errors. *Nature Neuroscience*, *13*(6), 669–71.

D'Argembeau, A., Collette, F., Van der Linden, M., Laureys, S., Del Fiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2005). Self-referential reflective activity and its relationship with rest: a PET study, *NeuroImage*, *25*(2), 616-624.

D'Argembeau, A., Jedidi, H., Balteau, E., Bahri, M., Phillips, C., & Salmon, E. (2011). Valuing one's self: medial prefrontal involvement in epistemic and emotive investments in self-views*. Cerebral Cortex*, *22*(3), 659-667.

de Greck, M., Rotte, R., Paus, D., Moritz, R., Thiemann, U., Proesch, U. Bruer, Moerth, S., Templemann, C., Bogerts, B., & Northoff, G. (2008). Is our Self based on reward? Self-relatedness recruits neural activity in the reward system. *Neuroimage*, *39*(4), 2066-2075.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., & Julie, A.F. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*(6), 3072-3077.

Den Ouden, H.E.M., Kok, P., & de Lange, F.P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19(8)*, 1153–1160.

Denny, B.T., Kober, H., Wager, T.D., & Ochsner, K.N. (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742–1752.

Enzi, B., de Greck, M. de, Prosch, U., Tempelmann, C., & Northoff, G. (2009). Is our Self nothing but reward? Neuronal overlap and distinction between reward and personal relevance and its relation to human personality. *PLoS ONE*, *4*(12), e8429-8429.

Epstein, S. (1973). The self-concept revisited: or a theory of a theory. *American Psychologist*, *28*(5), 404-404.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C.E., & Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science*, *318*(5854), 1305–1308.

Glimcher, P.W., Dorris, M.C., & Bayer, H.M. (2005). Physiological utility theory and the neuroeconomics of choice. *Games and Economic Behavior*, *52*(2), 213–256.

Glimcher, P.W., & Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. *Science*, *306*(5695), 447-452.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S.S., Maumet, C., & Margulies, D.S. (2015). NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, *9*(8).

Hampton, A.N., Bossaerts, P., & O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience*, *26*(32).

Henson, R. (2006). Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Sciences*, *10*(2), 64-69.

Kable, J.W. & Glimcher, P.W. (2009). The neurobiology of decision: consensus and controversy. *Neuron*, *63*(6), 733-745.

Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*(5), 785-794.

Krienen, F.M., Tu, P.C., & Buckner, R.L. (2010). Clan mentality: evidence that the medial prefrontal cortex responds to close others. *The Journal of Neuroscience*, *30*(41), 13906-13915.

Izuma, K., Saito, D.N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284–94.

James, W. (1890). The principles of psychology. New York: H. Holt & Company.

Jankowski, K.F., Moore, W.E., Merchant, J.S., Kahn, L.E., & Pfeifer, J.H. (2014). But do you think I'm cool? *Developmental Cognitive Neuroscience*, *8*, 40-54.

Kao, M.H., Madal, A., Lazar, N., & Stufken, J. (2009). Multi-objective optimal experimental designs for event-related fMRI studies. *Neuroimage*, *44*(3), 849-856.

Kamin, L.J. (1969). Selective association and conditioning – fundamental issues in instrumental learning (eds. Mackintosh, N.J. and Honig, W.K.) Dalhousie University Press, 42-64.

Kelly, W.M., Macrae, C.N., Wyland, C.L., Caglar, S. Inati, S., & Heatherton, T.F. (2002). Finding the self? An event related fMRI study. *Journal of Cognitive Neuroscience*, *5*(14), 785-794.

Klein, S.B., Loftus, J., & Burton, H.A. (1989). Two self-reference effects: the importance of distinguishing between self-descriptiveness judgments and autobiographical retrieval in self-referent encoding. *Journal of Personality and Social Psychology*, *56*, 853.

Krienen, F.M., Tu, P., & Buckner, R.L. (2010). Clan Mentality: Evidence That the Medial Prefrontal Cortex Responds to Close Others. *The Journal of Neuroscience*, *30*(41), 13906-13915.

Louie, K., Grattan, L., & Glimcher, P.W. (2011). Value-based gain control: divisive normalization in parietal cortex. *The Journal of Neuroscience*, *31*(29), 10627–10639.

Lipsman, N., Nakao, T. Kanayama, N., Krauss, J.K., Anderson, A., Giacobbe, P., Hamani, C., Hutchison, W.D., Dostrovsky, J.O., Andres, T.W., Lozano, M., & Northoff, G., (2014). Neural overlap between resting state and self-relevant activity in human subcallosal cingulate cortex – Single unit recording in an intracranial study. *Cortex*, *60*,139-144.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society: Biological Sciences*, *176*, 161-234.

Montague, P.R., & Berns, G.S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*(2), 265-284.

Montague, P.R. (2007). Neuroeconomics: a view from neuroscience. *Functional Neurology*, *22*(4), 219-234.

Montague, P.R., King-Casas, B., & Cohen, J.D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417–448.

Mitchell, J.P., Macrae, C.N., & Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655-663.

Moore, W.E., Merchant, J.S., Kahn, L.E., & Pfeifer, J.H. (2014). “Like me?” Ventromedial prefrontal cortex is sensitive to both personal relevance and self-similarity during social comparisons. *Social Cognitive and Affective Neuroscience*, *9*(4), 421-426.

Moran, J.M., Kelley, W.M., & Heatherton, T.F. (2013). What can the organization of the brain’s default mode network tell us about self-knowledge? *Frontiers in Human Neuroscience*, *7*, 391.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472.

Mussweiler, T., Rüter, K. & Epstude, K. (2004). The ups and downs of social comparison: mechanisms of assimilation and contrast. *Journal of Personality and Social Psychology*, *87*(6), 832-844.

Nicolle, A., Klein-Flugge, M., Hunt, L., Vlaev, I., Dolan, R., & Behrens, T. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron*, *75*(6), 1114–1121.

Northoff, G., Qin, P., & Feinberg, T. E. (2010). Brain imaging of the self-Conceptual, anatomical and methodological issues. *Consciousness and Cognition*, *20*(1), 52-63

Northoff, G. (2011). Self and brain: what is self-related processing? *Trends in Cognitive Sciences*, *15*(5), 186-187.

Northoff, G. & Hayes, D. J. (2011). Is our self nothing but reward? *Biological Psychiatry*, *69*(11), 1019–1025.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97-98.

Pfeifer, J.H., C.L. Masten, L.A. Borofsky, M. Dapretto, A.J. Fuligni, & M.D. Lieberman. (2009) Neural correlates of direct and reflected self-appraisals in adolescents and adults: when social perspective-taking informs self-perception. *Child Development*, *80*(4), 1016-1038.

Pfeifer, J.H. & Allen, N.B. (2012). Arrested development? Reconsidering dual-systems models of brain function in adolescence and disorders. *Trends in Cognitive Sciences*, *16*(6).

Pfeifer, J.H. & Peake, S.J. (2012). Self-development: integrating cognitive, socioemotional, and neuroimaging perspectives. *Developmental Cognitive Neuroscience*, *2*(1), 55–69.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59-63.

Poldrack, R.A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, *72*(5).

Potenza, M. (2008). The neurobiology of pathological gambling and drug addiction: an overview and new findings. *Philosophical Transactions of the Royal Society: Biological Sciences*, *363*(1507), 3181–3189.

Rangel, A., Camerer, C., & Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545-556.

Rescorla, R.A. & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. Classical Conditioning II: Current Research and Theory. (Eds Black, A.H., Prokasy, W.F.) New York: Appleton Century Crofts.

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Science*, *6*(3), 147-56.

Schacter, D.L., Addis, D.R., & Buckner, R.L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, *8*, 657-661.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Sutton, R. S. & Barto, A. G. (1998). Reinforcement Learning: An Introduction, MIT Press, Cambridge, Massachusetts.

Symons, C.S. & Johnson, B.T. (1997). The self-reference effect in memory: a meta-analysis. *Psychological Bulletin*, *121*(3), 371.

Tamir, D. I. & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences USA*, *107*(24), 10827-32.

Tamir, D. I. & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences USA*, *109*(21), 8038-8043.

Van Der Meer, L., Costafreda, S., Aleman, A., & David, A.S. (2010). Self-reflection and the brain: A theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. *Neuroscience and Biobehavioral Reviews*, *34*(6), 935-946.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3) 820-858.

Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E., & Gray, J.R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 470–7.

Whitfield-Gabrieli, S. & Ford, J.S. (2011). Default mode network activity and connectivity in psychopathology. *Annual Review of Clinical Psychology*, *8*, 49-76.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., & Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*, 665–670.