Neural Affective Decision Theory: Choices, Brains, and Emotions

Abninder Litt[1*], Chris Eliasmith[2, 3] and Paul Thagard[2, 4, 5]

ABSTRACT

We present a theory and neurocomputational model of how specific brain operations produce complex decision and preference phenomena, including those explored in prospect theory and decision affect theory. We propose that valuation and decision making are emotional processes, involving interacting brain areas that include two expectation-discrepancy subsystems: a dopamine-encoded system for positive events and a serotonin-encoded system for negative ones. The model provides a rigorous account of loss aversion and the shape of the value function from prospect theory. It also suggests multiple distinct neurological mechanisms by which information framing may affect choices, including ones involving anticipated pleasure. It further offers a neural basis for the interactions among affect, prior expectations and counterfactual comparisons explored in decision affect theory. Along with predicting the effects of particular brain disturbances and damage, the model suggests specific neurological explanations for individual differences observed in choice and valuation behaviors.

Keywords: computational neuroscience; decision making; emotion; framing; prospect theory.

[1] Graduate School of Business, Stanford University, Stanford, CA 94305-5015.
[2] Department of Philosophy. [3] Department of Systems Design Engineering. [4] Department of Psychology. [5] Cheriton School of Computer Science. University of Waterloo, Ontario, Canada, N2L 3G1.
[*]Correspondence concerning this article should be addressed to Abninder Litt, Graduate School of Business, Stanford University. E-mail: alitt@stanford.edu.

How do people decide what clothes to wear, what to eat for dinner, what car to buy, or what kind of career to pursue? In traditional economics, the standard answer is that people decide by maximizing expected utility, but psychologists have found many problems with this kind of decision theory as a description of human behavior (e.g., Camerer, 2000; Kahneman & Tversky, 2000; Koehler, Brenner, & Tversky, 1997; Rottenstreich & Hsee, 2001; Tversky & Kahneman, 1991). Economists commonly take preferences as given, but from a psychological point of view it should be possible to explain how preferences arise from cognitive and affective processes. Work in this spirit has made tremendous progress in revealing key features and dynamics missed by theories disconnected from the study of cognitive, emotional and socially motivated phenomena, such as a common hypersensitivity to losses over equivalent gains (Kahneman & Tversky, 1979) and the affective influence of prior expectations and counterfactual comparisons on preference judgments (Mellers, 2000). Moreover, with the rise of cognitive and affective neuroscience, it should be possible to identify precise neural mechanisms underlying these behavioral-level explanations of why people make the choices that they do.

We propose *neural affective decision theory* as a psychologically and neurologically realistic account of specific brain mechanisms underlying human preference and decision. The theory consists of four principles, which we shall list here and describe in detail later:

1. Affect. Decision making is a cognitive-affective process, crucially dependent on emotional evaluation of potential actions.

2. Brain. Decision making is a neural process driven by coordinated dynamic interactions among multiple brain areas, including parts of prefrontal cortex as well as major subcortical systems.

3. Valuation. The brain forms preferences via interacting but distinct mechanisms for positive and negative outcomes, encoded primarily by dopamine and serotonin, respectively.

4. Framing. Judgments and decisions vary depending on how the context and manner of the presentation of information initiate different neural activation patterns.

There is substantial empirical evidence for each of these principles, and when integrated in the precise manner we outline they can explain the findings of a wide range of psychological and neurological phenomena.

In order to connect these principles with experimental results in a mathematically and neurologically rigorous fashion, we have developed a neurocomputational model called ANDREA (Affective Neuroscience of Decision through Reward-based Evaluation of Alternatives). It operates within the Neural Engineering Framework (NEF) developed by Eliasmith and Anderson (2003), using biologically realistic populations of neurons to encode and transform complex representations of relevant information. ANDREA simulates computations among several thousand neurons to model coordinated activities in seven major brain areas that contribute to valuation and decision making: the amygdala, orbitofrontal cortex, anterior cingulate cortex, dorsolateral prefrontal cortex, the ventral striatum, midbrain dopaminergic neurons, and serotonergic neurons centered in the dorsal raphe nucleus of the brainstem.

ANDREA successfully produces detailed neural-level simulations of behavioral findings explored in prospect theory (Kahneman & Tversky, 1979) and the decision affect theory of Mellers and colleagues (1997). It shows how specific neural processes can produce behaviors observed in both psychological experiments and real-world scenarios that have provided compelling evidence for these preference and choice theories. In particular, ANDREA provides neurological explanations for the major hypothesis of prospect theory that losses have greater psychological force than gains, as well as for the fundamental claim of decision affect theory that the evaluation (and subsequent potential choice) of an option is strongly influenced by its

perceived *relative pleasure*, an emotional determinant that is dependant on expectations and counterfactual comparisons. In our concluding discussion, we compare ANDREA to other models in decision neuroscience, describe promising avenues of expansion for ANDREA and neural affective decision theory, and suggest additional psychological phenomena that are likely to fall within the scope of our theory.

NEURAL AFFECTIVE DECISION THEORY

We now examine in detail the four guiding principles of neural affective decision theory, including connections to and supporting evidence provided by a diverse array of research in both psychology and neuroscience. The ANDREA implementation of the theory we describe later provides the formal integration of these ideas necessary for our detailed simulation experiments.

Principle 1. Affect.

According to our first principle, decision making is a cognitive-affective process, crucially dependent on emotional evaluation of potential actions. This claim rejects the assumption of traditional mathematical decision theory that choice is a 'cold' process involving the calculation of expected values and utilities (Kreps, 1990; von Neumann and Morgenstern, 1947). The original nineteenth-century concept of utility was a psychologically rich, affective one based on pleasure and pain (Kahneman, Wakker, and Sarin, 1997). In contrast, twentieth-century economics adopted the behaviorist view that utilities are mathematical constructions based on preferences revealed purely by behavior. There is no room in this view for findings observed in both psychological experiments and everyday life that people's decisions are often highly emotional, with preferences arising from having positive feelings for some options and negative ones for others. While psychology has introduced a more complex characterization of the cognitive processes underlying decision making, the specific influence of affect on behavior

has frequently been ignored. Rottenstreich and Shu (2004) argue that this neglect of affect may stem from an original desire of psychological decision researchers to minimize differences with the terminology and general themes of classical normative decision theories.

But there is increasing appreciation in cognitive science that emotions are an integral part of decision making (e.g. Bechara, Damasio, & Damasio, 2000, 2003; Churchland, 1996; Lerner & Keltner, 2000; Loewenstein, Weber, Hsee, & Welch, 2001; Sanfey, Loewenstein, McClure, & Cohen, 2006; Slovic, 2002; Wagar & Thagard, 2004). Kahneman (2003, p. 710) argues that "there is compelling evidence for the proposition that every stimulus evokes an affective evaluation." Common experience suggests that emotions are both inputs and outputs of decision making. Preference for one option over another depends strongly on their relative emotional interpretations, and the process of decision making can itself generate emotions such as anxiety or relief. The relevance of emotion to decision making is consistent with physiological theories that regard emotions as reactions to somatic changes (James, 1894; Damasio, 1994). It also fits with some cognitive theories of emotions, which regard them as judgments about the extent to which ones goals are being satisfied (Oatley, 1992). From a neurological perspective, it is easy to see how emotions can be both cognitive and physiological, as there are numerous interconnections among the relevant brain areas.

Principle 2. Brains.

According to our second principle, decision making is a neural process driven by coordinated dynamic interactions among multiple brain areas, including parts of prefrontal cortex as well as major subcortical systems. In particular, activity in brain regions involved in assessing and acting upon the appetitive or aversive nature of stimuli (commonly conceptualized as part of the brain's *reward system*) seems most crucial to understanding judgment and choice behavior

(for a review, see Sanfey, Loewenstein, McClure, & Cohen, 2006). Empirical neuroscientific investigation of the nature of preference and decision has been developing rapidly, and much work today is identifying specific brain areas involved in producing decision-related behaviors (e.g., Bayer & Glimcher, 2005; Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; Knutson et al, 2005; McClure, York & Montague, 2004; Montague & Berns, 2002). This nascent field of *decision neuroscience* represents an exciting frontier of deep exploration into how and why people act, think and feel as they do in choice and judgment scenarios (Shiv et al., 2005).

But taking a neural approach to decision making allows for much more than simply identifying brain areas activated in the subjects of behavioral studies. The development of biologically plausible theories of *how* brain areas interact to produce preferences and choices can provide more refined mechanistic explanations of decision behaviors. Moreover, investigation at the neural level can suggest novel experiments, for example the recent discovery that an odorless nasal spray preparation of the neuropeptide oxytocin increases trust in risky choice scenarios, including those involving monetary transactions (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). Such a finding is one that decision neuroscience can reveal, but that would be missed by higher level psychological study alone. Neuroscience can thus inform the development of more detailed predictions and richer understandings of behavioral-level observations.

Principle 3. Valuation.

Our third principle states that the brain forms preferences via interacting but distinct mechanisms for positive and negative outcomes, encoded primarily by dopamine and serotonin, respectively. There is extensive evidence that midbrain dopamine neurons, such as those in the ventral tegmental area and nucleus accumbens, are involved in the computation of a discrepancy between the expected and actual rewarding nature of an outcome (e.g., Schultz 1998, 2000; Suri

2002; Knutson et al. 2005), although recent evidence suggests that this activity is only involved in the encoding of *positive* deviations from expectations, that is, getting more than one expected. (Bayer & Glimcher, 2005). Daw, Kakade, and Dayan (2002) describe a plausible alternative brain mechanism for situations in which one receives less than expected, arguing that serotonin innervation from the dorsal raphe nucleus of the brainstem is crucial for producing characteristic reactions to negatively valued stimuli and matters being considered (e.g., options in a choice scenario). There are thus neurobiological reasons for viewing gains and losses as being encoded and subsequently assessed in a fundamentally different manner by the brain, involving distinct neural circuits and activation patterns. This provides the basis for our explanation of the central finding of prospect theory that losses loom larger than gains. Our neurocomputational model ANDREA simulates how interactions of the dopamine and serotonin systems with the amygdala and other brain areas may enable this asymmetric assessment of positive and negative outcomes.

Principle 4. Framing.

The last principle states that judgments and decisions vary depending on how the context and manner of the presentation of information initiate different neural activation patterns. The importance of framing is evident from the long history of influential work by Kahneman and Tversky (1981, 1986, 2000). They demonstrated that framing a decision in terms of either losses or gains can substantially affect the choices that people make, and related phenomena have been observed in many real-life arenas such as the stock market and consumer choice (Camerer, 2000). We contend that framing can be understood even more deeply from the neural-affective perspective we propose in our first two principles. The simulation results we describe later show how this enriched conception of framing allows for the integration of diverse lines of behavioral decision research, as well as the postulation of important new predictions and hypotheses.

We take the concept of framing to encompass any potential effects of the manner or context of presentation on decisions and judgments. Following this characterization, such findings as preference reversals when outcomes are evaluated jointly versus separately (e.g., Hsee, Loewenstein, Blount, & Bazerman, 1999) might also be considered to be framing results. Another such framing effect is illustrated by the 'trolley-footbridge' dilemma (e.g., Greene et al., 2001): most people consider flipping a switch to kill one person instead of five morally justified, but consider it immoral to personally push a person into the path of an oncoming trolley, killing that person but preventing the trolley from killing five others. We will also explain some of the findings of the decision affect theory of Mellers and colleagues (1997) as framing effects that differentially activate specific neural systems.

These four principles make strong claims about the processes that constitute human decision making, but alone they are not sufficiently precise to explain particular experimental results. We now describe a rigorously defined, biologically realistic neurocomputational model that specifies how different brain areas might interact in a manner consistent with neural affective decision theory to produce observed behavioral phenomena.

THE ANDREA MODEL

A neuropsychological theory consists of a set of hypotheses about how specific brain operations produce observed behaviors. Because of the complexity of the brain, computational models are indispensable for theoretical neuroscience, both for precisely specifying relevant neural structures and activities and for examining their implications through appropriate simulation experiments. Litt, Eliasmith, and Thagard (2006) proposed a biologically detailed neural model of reward system substrates of valuation and decision. We describe here the primary functional components of this model, and introduce several explanatorily valuable

additions regarding neural response characteristics and the complexity of emotional arousal encoding. This version of the model we call ANDREA, for Affective Neuroscience of Decision through Reward-based Evaluation of Alternatives.

Our model applies the Neural Engineering Framework (NEF) developed by Eliasmith and Anderson (2003), and has been implemented in MATLAB using the NEF simulation software *NESim* (see Appendix A). Neural populations ('ensembles') and their firing activities are described in the NEF in terms of mathematically precise representations and transformations, with the dynamic characteristics of neural computations characterized using the tools of modern control theory. Appendix B outlines the exact mathematical nature of representation, transformation and dynamics as defined by the NEF. This rigorous, generalized mapping of high-level mathematical entities and transformations onto biophysical phenomena such as spike patterns and currents allows for biologically constrained computations and dynamics to be implemented in physiologically realistic neural populations, and has proven successful in modeling phenomena ranging from the swimming of lamprey fish (Eliasmith & Anderson, 2000) to the Wason card task from cognitive psychology (Eliasmith, 2005b).

Figure 1 shows the connectivity structure between the different brain regions we have modeled. A comprehensive examination of afferent and efferent transmission among these regions would feature many more connections than we have included. The interactions shown represent particular paths of coordinated activity that contribute to observed behaviors, rather than a full characterization of all relevant neural activity. Appendix A provides details regarding the specific numbers of neurons used to model each of these brain areas, as well as the physiological parameters used to model individual neurons in each of these populations. Each input-output relation symbolized by a connection line in Figure 1 maps onto one or more specific

mathematical transformations, as summarized in Appendix C. We now describe these coordinated neural computations as they are relevant to explaining decisions and valuations.

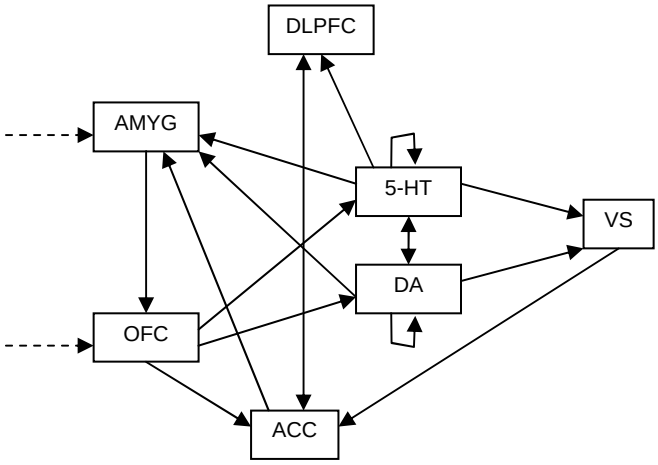

Figure 1: Basic connectivity framework. Dotted arrows represent external inputs to the model. Abbreviations: 5-HT, dorsal raphe serotonergic neurons; ACC, anterior cingulate cortex; AMYG, amygdala; DA, midbrain dopaminergic neurons; DLPFC, dorsolateral prefrontal cortex; OFC, orbitofrontal cortex; VS, ventral striatum.

Subjective Valuation by Emotional Modulation

Valuation of alternatives and other information is an essential part of decision making. Central to the performance of this task by ANDREA is an interaction between the amygdala and orbitofrontal cortex (Fig. 1). Much research has implicated orbitofrontal cortex in the valuation of stimuli (e.g., Rolls, 2000; Thorpe, Rolls & Maddison, 1983), particularly in light of its extensive connections with sensory processing areas of the brain. Several recent studies have indicated an important role for orbitofrontal neurons in providing a sort of "common neural currency" (Montague & Berns, 2002) which allows for the evaluation and comparison of figurative (or even literal) apples and oranges (Padoa-Schioppa & Assad, 2006). Recent studies of the amygdala have challenged its traditional association with mainly aversive stimulus processing, showing instead activation based on the degree to which stimuli are salient or arousing, rather than a specific valence type (for a review, see McClure, York & Montague,

2004). This has inspired a reinterpretation of classic results as indicating that negatively appraised events may be in general more emotionally arousing than positive outcomes, perhaps because of a need to alter current behavior in response to aversive feedback. In accord with research on the role of the amygdala in emotional attention (Adolphs et al., 2005) and multiplicative scaling observations for visual attention (e.g., Treue, 2001), ANDREA performs a multiplicative modulation by amygdala-encoded emotional arousal of the valuation computation performed in orbitofrontal cortex (Fig. 2). That is, orbitofrontal valuations are modeled as being multiplicatively dampened or intensified, depending on whether the individual is in a lowered or heightened state of affective arousal, respectively.

Let V represent baseline orbitofrontal stimulus valuation, based on initial sensory and cognitive processing and provided as an input in our model. Taking A to represent amygdala-encoded emotional arousal, we characterize the output *subjective* valuation S at time t as:

$$S(t) = A(t) \cdot V(t). \qquad [1]$$

Thus, increased levels of emotional arousal will amplify the subjective valuation of stimuli by orbitofrontal cortex, while lower arousal levels lead to valuation attenuation. As we discuss in our later account of prospect theory, ANDREA also introduces realistic biological constraints imposed by neural firing saturation that help to explain valuation behaviors observed in humans, an advancement over our earlier reward system model (Litt, Eliasmith, & Thagard, 2006).

Surprise as Deviation from Expectations

Figure 2 shows that our model generates amygdala activity upsurges accompanying changes in the valuation input to orbitofrontal cortex, and that these upsurges are valence-asymmetric: negative changes in valuation (losses) produce greater affective arousal increases than equivalent positive changes (gains). This neurological behavior is produced mainly through

a modulation of amygdala-encoded emotional arousal by a *reward prediction error* signal. This is simply the discrepancy between expected and actual stimulus valuation, and as such represents the effect of the surprising nature of a stimulus on how emotionally arousing it is.

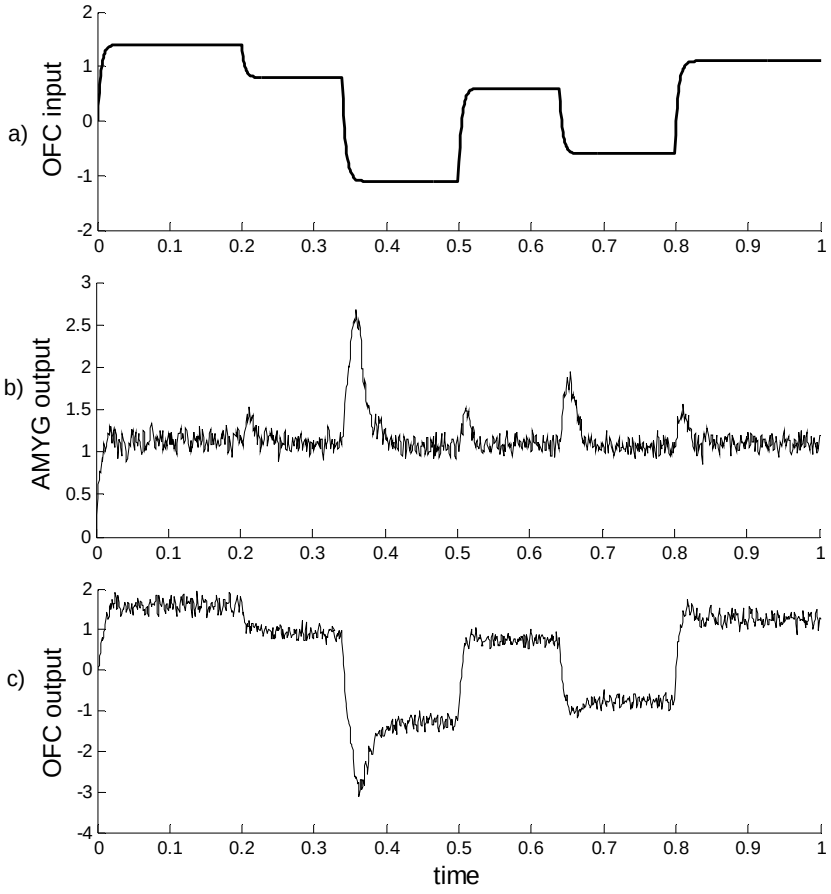

Figure 2: Arousal modulation of valuation. **a)** The "emotionless" input signal to orbitofrontal cortex consists here of positive and negative valuation changes of varying magnitude. The vertical axis can be interpreted as a sort of neural currency scale: upward steps in the graph thus represent gains, while stepping down indicates a loss. Positive/negative sign corresponds to appetitive/aversive valence. **b)** Emotional arousal reflected in amygdala activity. Decoded output from spiking neuron populations (see Appendix B). Upsurges correspond to arousal increases coinciding with changes in the externally provided stimulus value signal in a), demonstrating the role such changes play in influencing emotional engagement. **c)** Multiplicative modulation of the activity presented in a) by that shown in b). Emotional arousal can induce significant changes in valuation from the baseline input signal.

For this computation we employ the temporal difference (TD) model (Sutton & Barto, 1998), due to its simple mathematical structure and robust correspondence with experimental

neural activity observations (e.g., Schultz, 2000). TD computes reward prediction error (*E*) based on the difference between the latest reward valuation and a weighted sum of all previous rewards (*P*). Using our arousal-modulated signal *S* as the input regarding current stimulus valuation, this leads to the modeled recurrent equations

$$E(t) = S(t) - P(t - 1) \quad [2]$$

$$P(t) = P(t - 1) + \alpha \cdot E(t), \quad [3]$$

where α is a learning rate constant between 0 and 1. This activity has typically been modeled by increased midbrain dopamine firing with positive prediction errors (that is, getting more than expected) and firing rate depression for negative errors (getting less than expected) (Schultz, 1998, 2000; Suri, 2002).While this approach seems valid for particular ranges, recent work has called into question the feasibility of midbrain dopamine acting alone to perform this computation (Daw, Kakade, & Dayan, 2002; Dayan & Balleine, 2002). In particular, such physiological constraints as low baseline firing rates make it difficult to envision how activity depression could be used to well encode highly negative prediction errors, a concern supported by recent experimental findings (Bayer & Glimcher, 2005).

Accordingly, we adopt an interacting opponent encoding of positive prediction errors by midbrain dopamine and negative errors by serotonergic neurons in the dorsal raphe nucleus of the brainstem. This is supported by a variety of experimental studies in humans and other animals (Deakin, 1983; Evenden & Ryan, 1996; Mobini et al., 2000; Soubrié, 1986; Vertes, 1991; for a review, see Daw, Kakade, & Dayan, 2002). By separating the encodings of losses and gains, we are able to *distinctly calibrate* the modulatory effects of positive and negative valuation changes to provide a plausible neural mechanism for loss aversion (Appendix C). That is, asymmetries in loss-gain valuation can be modeled via differing amygdala sensitivity to

inputs from dorsal raphe or midbrain areas, perhaps realized in actual brains through differences in specific neurotransmitter receptor concentrations in the amygdala, or similar mechanisms of connectivity strength variation (e.g., receptor sensitivity differences).

Increased Behavioral Saliency of Negative Outcomes

Valence-asymmetry in emotional arousal is further strengthened in our model through the activities we have assigned to the anterior cingulate and dorsolateral prefrontal cortices, specifically via a proposed dissimilarity in the influences of losses and gains on required behavioral planning. Much evidence supports the importance of dorsolateral prefrontal cortex in the planning, representation and selection of goal-directed behaviors (e.g., Owen, 1997), and the anterior cingulate cortex in the detection of conflicts between current behavior and desired or expected results, interfacing appropriately with dorsolateral prefrontal in the process (e.g., Bush, Luu, & Posner, 2002). We hypothesize an increased behavioral saliency of *negative* reward prediction errors, as such results may indicate that current behavior needs to be modified, rather than be simply maintained or strengthened as a positive error would indicate. Such a situation would introduce the attendant cognitive resource requirements of new action plan formation and execution in response to the displeasing outcome, as well as potential environmental risks stemming from altering current behavior. We model this increased behavioral saliency through a corresponding increase in emotional arousal (Appendix C). Thus, feedback from dorsolateral prefrontal cortex and the anterior cingulate to the amygdala in our model further increases the affective influence of losses over similarly sized gains.

In combination with the previously discussed roles of dopamine and serotonin in influencing the amygdala, we arrive at our final characterization of how emotional arousal A is

influenced by the saliency of unexpected gains and losses, as well as potential behavioral modification costs associated with the latter:

$$A(t) = A_1(t) + \beta \cdot DA(t) + \gamma \cdot 5\text{-}HT(t) + C(t). \quad [4]$$

A_1 represents a degree of emotional arousal determined by external factors unrelated to reward prediction error or the described dorsolateral-cingulate contribution. In previous work we have provided this as a straightforward input signal to the model (Litt, Eliasmith, & Thagard, 2006). We shall describe later how ANDREA expands A_1 arousal by incorporating prior expectations regarding valuation targets, which allows for a neurobiological explanation of decision affect theory (Mellers, Schwartz, Ho, & Ritov, 1997). *DA* and *5-HT* are the opponent encodings of positive and negative reward prediction error, respectively, with γ chosen to be a connection-strength constant greater than β to simulate an increased influence of serotonin-encoded losses over dopamine-encoded gains on emotional state. Finally, *C* represents the additional costs associated with losses that increase their behavioral saliency, and hence emotional import, as determined by activity in the anterior cingulate and dorsolateral prefrontal cortices.

A NEURAL ACCOUNT OF PROSPECT THEORY

Prospect theory, a theoretical framework for understanding risky choice developed by Kahneman and Tversky (1979, 1982, 2000), has been applied to many preference and choice behaviors commonly exhibited by people. The most famous such phenomenon is loss aversion, whereby people behave asymmetrically in their personal valuations of objectively equivalent losses and gains. Central to prospect theory's resolution of this and other inconsistencies between classical decision research and actual behavior is a redefined characterization of the nature of subjective evaluations of decision outcomes. The resulting *value function* proposed by the theory has the following essential characteristics (Kahneman & Tversky, 1979; Tversky & Kahneman,

1984): i) subjective value is defined on gains and losses—that is, deviations from a neutral reference point—rather than on total wealth, as is typical of expected utility theory; ii) the value function is concave for gains and convex for losses; and iii) the curve is steeper for losses than gains. Taken together, an asymmetric sigmoid value function is the well-known result (Fig. 3).

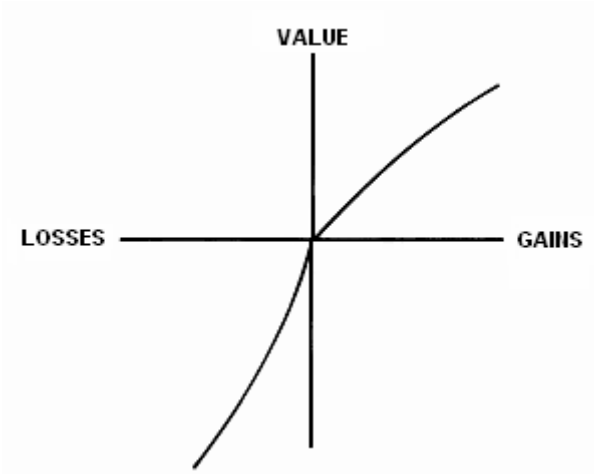

Figure 3: A hypothetical prospect theory value function, illustrating subjective valuation commonalities observed in tests of numerous subjects.

Loss Aversion

Neural affective decision theory, via the ANDREA model, provides a compelling explanation of loss aversion. The combination of arousal modulation of subjective valuation and the increased affective import of losses produces emotionally-influenced orbitofrontal valuations that overweight losses. This can be seen in Figure 2, and even more vividly in the simplified simulation of Figure 4. The resulting effects on thinking and behavior would produce the asymmetries in peoples' evaluations of and responses to gains and losses that have been documented and explained by prospect theory.

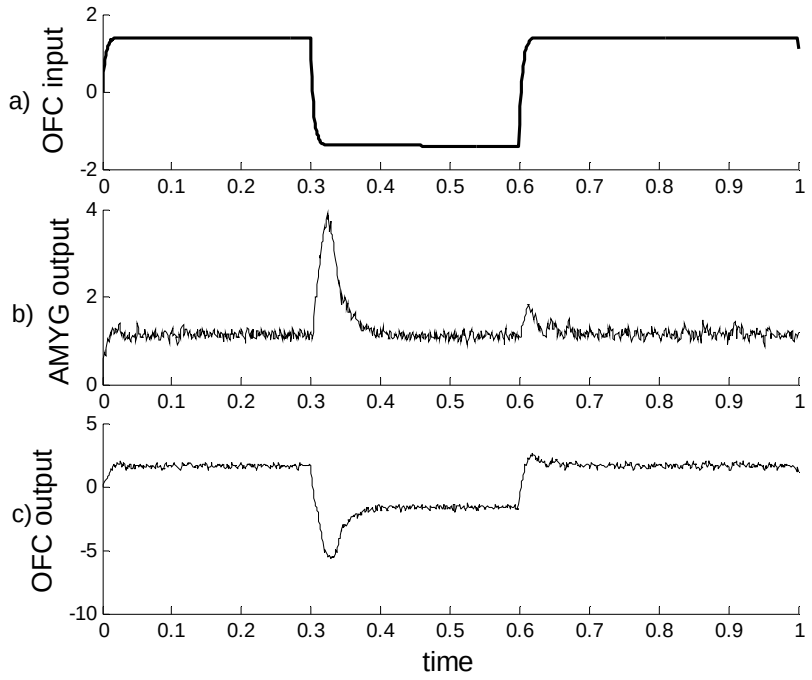

Figure 4: Unbalanced evaluation of gains and losses. **a)** The input signal to orbitofrontal cortex consists of positive and negative changes in value of equal magnitude. **b)** Arousal level modulated by prediction error and likely behavioral saliency of stimuli. The loss induces a much greater arousal increase than the equal gain. **c)** The outcome of the unevenness displayed in b). Reductions in stimulus valuation (losses) are disproportionately amplified compared to gains.

We turn now to extending this neural account of loss aversion into a detailed biological explanation for the specific shape of prospect theory's sigmoid value function.

The Value Function of Prospect Theory

In developing a neural theory of preference that meshes with the behaviorally inspired prospect theory value function, the first step is to identify brain regions that should be expected to produce responses corresponding to the sorts of behaviors monitored in psychological studies of preference and valuation. As discussed earlier, it seems natural to look to orbitofrontal cortex as the site of activity mapping directly onto people's subjective valuations of gains and losses. It has been implicated strongly in tasks related to valuation and comparison of outcomes, events and perceived stimuli in general, and we have described a fundamental affective modulation of this encoding that may form the basis of the subjective nature of ultimate outcome valuation.

The next step is to identify features of the ANDREA model that might explain the specific nature of the sigmoid value function, as described previously in terms of three primary characteristics (Fig. 3). Feature i) of the curve, valuation in terms of reference point deviation, identifies the sort of input signal to be modulated in orbitofrontal cortex. Since the degree to which a stimulus is considered a loss or a gain is a representation of its divergence from a neutral reference, evaluating such changes in value calls for a step-style input, similar to those shown in Figures 2 and 4, where the subjective valuation of a deviation of size *X* will be determined by the emotional modulation produced by an input valuation step from 0 to *X*. For example, to produce orbitofrontal activity corresponding to the subjective valuation of a loss of $200, we measure the emotionally modulated output from orbitofrontal cortex to a step input that moves from 0 (the reference point) to our target value (-200). Leaving aside feature ii) for a moment, the third aspect of the prospect theory value curve, a steeper slope for losses than gains, is simply loss aversion (Kahneman & Tversky, 1979). The biological account of loss aversion we previously described will thus serve as a critical component of our neural explanation of the S-curve.

The second feature of the sigmoid value function, the leveling off of loss and gain valuations at the extremes, requires appeal to additional neurological mechanisms. In particular, we introduce the notion of *neural saturation*. Any type of neuron has hard biological constraints on how fast it can fire. Each action potential is followed by a refractory period of repolarization during which the neuron cannot fire, and issues such as cellular respiration requirements and local neurotransmitter depletion introduce unavoidable limitations on spike rates. In the context of neurocomputation in the NEF (and hence ANDREA), this means that the range of values that can be encoded by any neural population is limited. This inherent restriction can actually serve an explanatory purpose in the case of the prospect theory value function. To explain how

ANDREA could produce leveling-off valuations at the extremes, note that we encode values through increasing neural spike rates in the encoding population as these values become larger (in either the positive or negative direction). Thus, in attempting to provide subjective valuations for very large gains or losses, neurons in orbitofrontal cortex will begin to saturate, as they simply cannot fire fast enough to produce linearly distinctive affective responses to increasingly large value deviations.

An alternative encoding of value magnitude through activated population *size*, rather than the firing levels of a fixed population, would seem to deny this saturation-based account of diminishing marginal sensitivity. Such a scheme might be akin to accumulator models that have been applied to numerosity encoding in intraparietal regions (Roitman, Brannon, & Platt, 2007). However, applying this approach to valuation encoding seems less plausible from a neurophysiological resources perspective; while the salient brain areas do have millions of neurons, all of these would require direct connectivity to areas interpreting magnitude information in order to impart the same information as naturally rate-tuned transmitter release by a population which encodes magnitude via firing rates. Indeed, accumulator models generally allow for cardinal value encodings on restricted integer scales such as 2-to-32, rather than the potentially arbitrary and quasi-analog scale on which valuations often lie.

Figure 5 illustrates the results of running simulations based on the preceding description. Each data point at *X* along the horizontal axes of 5b and 5c is the result of measuring a modulated orbitofrontal valuation output in a simulation providing orbitofrontal cortex a step input from 0 to *X*, in accord with the reference-deviation characterization of value in prospect theory. As expected, the effects of loss aversion are clear, with the slope differential indicating a greater affective impact of losses over equivalent gains. The 2:1 slope ratio observed here for

moderate losses and gains mirrors behavioral evidence in the decision literature (e.g., Kahneman, Knetsch, & Thaler, 1991). Finally, the concavity features of the value function have been successfully replicated in these simulations. Figure 6 shows the specific role of neural saturation in this regard. Each row in these spike rasters represents an individual orbitofrontal cortex neuron, and each point represents a single action potential at a specific point in time produced by the neuron in question. Clearly, equal changes in the size of a loss or gain do not necessarily produce similar changes in neural spiking, particularly as neurons begin to saturate. This causes a decreasing distinctness in firing response at the extremes, which we propose as the neurological basis for the leveling-off in loss and gain valuation observed in behavioral studies. Overall, the mechanisms we have outlined combine to produce a detailed, biologically plausible neural explanation of the nature of the value function described by prospect theory.

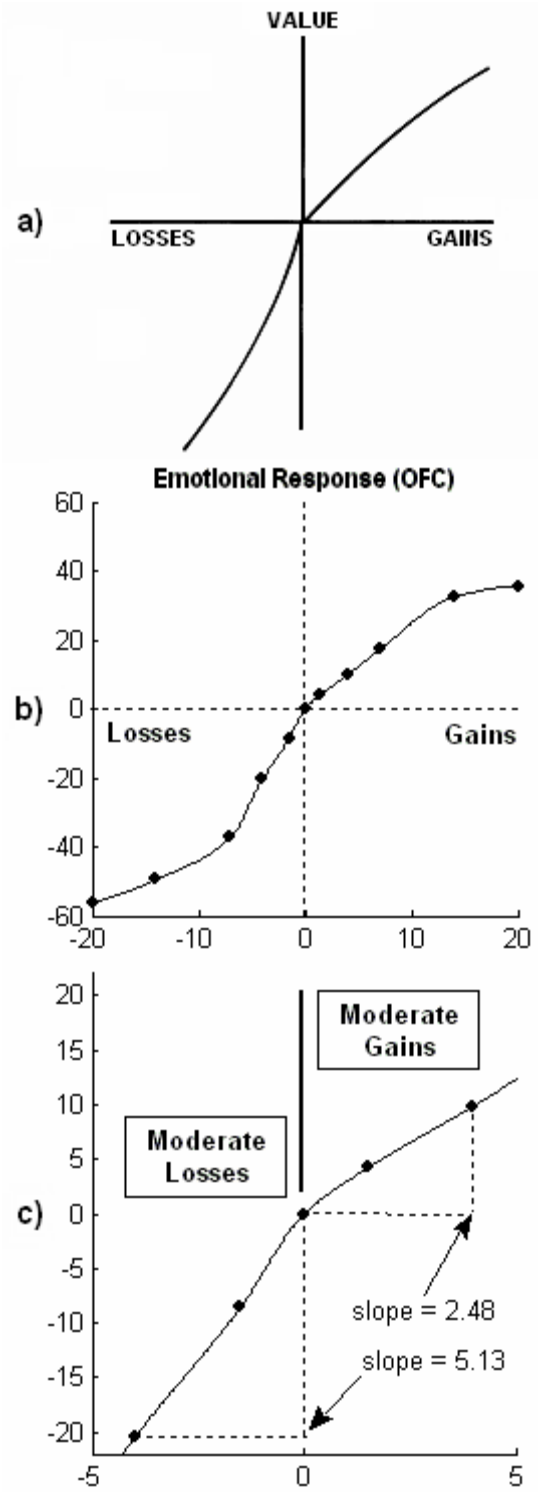

Figure 5: Value function simulation results. **a)** A typical prospect theory value function. **b)** Subjective valuation outputs from orbitofrontal cortex. Each data point represents the emotionally modulated valuation of the loss or gain value chosen along the horizontal axis. **c)** Close-up of the central portion of b), showing the differing slopes for loss and gain valuations.

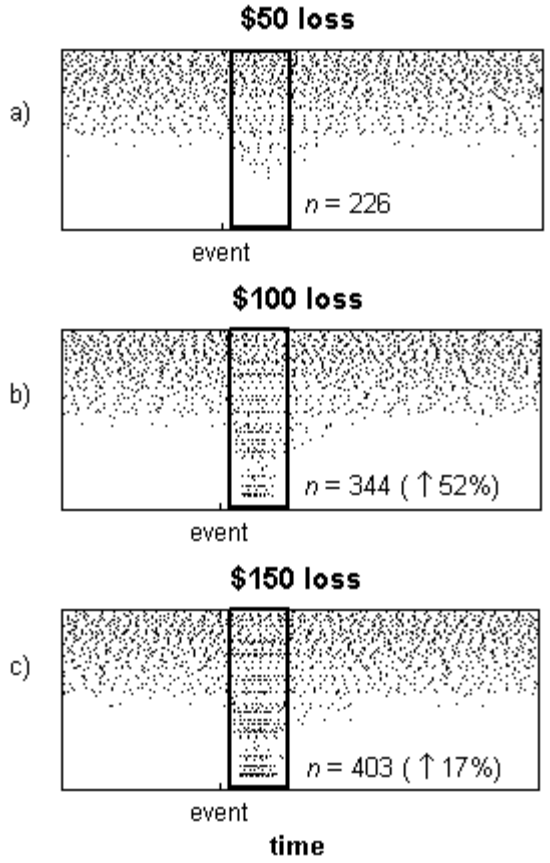

Figure 6: Orbitofrontal cortex spike rasters. Note the clear difference in activity between a) and b) (a 52% increase in immediate post-event spiking). This is in contrast to the relative similarity in spiking in b) and c) forced by neural saturation, as the $50 change moves farther away from the reference value of 0. This response characteristic forms the basis of our neural explanation for the concavity features of the prospect theory value function.

Framing through Reference Value Manipulation

Understanding how individuals respond differently depending on the manner in which information regarding a situation is presented is considered to be one of the primary explanatory successes of prospect theory. A famous illustration of the power of framing is the presentation of two different choice-sets to subjects regarding the consequences of different plans to handle the outbreak of a disease expected to kill 600 people (Tversky & Kahneman, 1981, 1986):

Problem 1: Program A – 200 people will be saved.

Program B – 1/3 probability 600 are saved, 2/3 probability nobody is saved.

Problem 2: Program C – 400 people will die.

Program D – 1/3 probability nobody will die, 2/3 probability 600 will die.

Faced with the choice in Problem 1, 72 percent of subjects chose Program A over B, whereas only 22 percent of subjects chose Program C over D when faced with Problem 2. Clearly, though, Programs A and C are objectively equivalent, as are their respective alternatives. The framing of situations in terms of losses or gains may thus cause dramatic reversals of preference in decision scenarios.

The mechanisms implemented in the ANDREA model provide a realistic neural basis for such framing effects. We describe means by which objectively equivalent outcomes can produce markedly different subjective valuations in orbitofrontal cortex, depending on the manner in which each is framed. In particular, note that feature i) of the prospect theory value function, the definition of subjective valuation on deviations from some neutral reference value, points towards an obvious mechanism for the framing of decisions: *variation of the reference value itself*. In the disease example, Problem 1 is framed in terms of lives saved rather than lives lost, while the reverse is true for Problem 2. Thus, choosing "zero" reference points from which subjective valuations shall deviate in each case should lead to different values for each problem. For Problem 1, "zero lives saved" would indeed correspond to 0 on a scale measuring the total number of people of our original 600 who are expected to be left alive after the choice of a given program for combating the disease. Crucially, however, the "zero lives lost" reference point for Problem 2 would correspond to the value *600* when measured on this same scale, since 600 out of 600 people alive indicates that no lives have been lost, as required.

Thus, in the case of the Program A option in Problem 1, a subjective valuation deviation described as "200 people out of 600 will be saved" represents a *positive-direction* deviation from

0 lives saved to 200 lives saved. In contrast, the objectively equivalent Program C option in Problem 2, described as “400 people out of 600 will die”, represents a *negative-direction* deviation from 0 lives lost to 400 lives lost, that is, from 600 left alive down to 200 left alive. Note that both deviation construals end at the value 200 on the scale of people still alive, since they are objectively equivalent outcomes. Nevertheless, because of our multiplicative modulation of valuation deviations by emotional arousal, opposite directions of deviation will produce subjective valuations that are emotionally amplified in opposite directions. Figure 7 illustrates the results of characterizing this type of framing as a manipulation of the deviation reference point. We obtain a subjective valuation of orbitofrontal step-input corresponding to Program A that is much more positive than that of a step-input corresponding to Program C, simply because of opposite directions of emotional modulation. This would explain the preference reversal that occurs upon switching decision frames, as what was seen as a gain in comparison to one reference value is suddenly evaluated as a loss in comparison to a different referent. Later we will discuss framing effects that operate in ways other than varying reference values, and how these different sorts of framing can in combination explain the observed interactions between affect, prior expectations and counterfactual comparisons explored in decision affect theory.

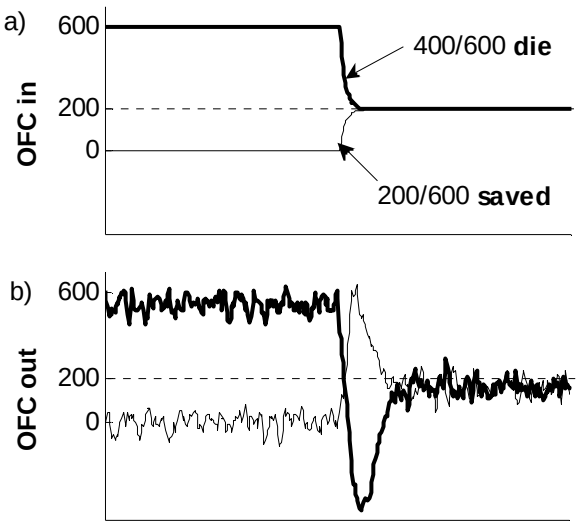

Figure 7: Simulation results for framing in terms of gains and losses. **a)** Objectively equivalent outcomes (ending up with 200 people alive) evaluated as deviations from different reference points. The thin-lined step input represents Problem 1/Program A, and the heavy line Problem 2/Program C (Tversky & Kahneman, 1981). **b)** Opposite directions of deviation produce opposite directions of emotional amplification in subjective valuation, leading to more a positive outlook towards Program A than Program C.

Predictions

Our neurological explanation of prospect theory suggests a range of testable neural-level predictions and hypotheses. Litt, Eliasmith and Thagard (2006) outline several such predictions in relation to loss aversion and the behavioral influence of serotonin. For example, the extent of a particular individual's hypersensitivity to losses is hypothesized to be correlated with the concentration in the amygdala of a specific serotonin receptor subtype, which would influence the degree to which negative reward prediction errors affect amygdala activity. As well, degraded connectivity between midbrain dopamine neurons and the raphe serotonin system is predicted to increase emotionally influenced overvaluation of *both* gains and losses, due to mutual attenuation effects that we have modeled between these systems. Such correlation between loss and gain sensitivity has indeed been shown in recent work .by Tom and colleagues (2007). The particular neural activity they describe suggests the mechanism underlying this

relationship may involve important additions to the computations captured in the ANDREA model, such as the effects of noradrenergic circuits.

Further empirical investigation of the neural correlates of loss aversion can provide more such tests and potential falsifications of the relationships proposed in ANDREA, although practical barriers to imaging brain stem serotonergic activity limit the capacity of the approach of the Tom et al. study in this respect. Additionally, this work and other explorations of "prospect theory on the brain" by this team did not show any significant amygdala activity relevant to loss aversion (Tom, Fox, Trepel, & Poldrack, 2007; Trepel, Fox, & Poldrack, 2005). Besides the studies we have previously cited that indicate an important role for the amygdala in emotional valuation, significant amygdala activity directly corresponding to loss aversion has contrastingly been directly observed in other fMRI experiments (e.g., Weber et al., 2007). Such discrepancies and limitations might be resolved by employing alternative experimental techniques; for instance, it is possible to temporarily diminish cerebral serotonin levels by an ingestion of a tryptophan depleting drink (Cools et al., 2005), which our model suggests would specifically diminish loss aversion. Depletion studies of decision-related phenomena hold much promise for testing the validity of neurocomputational models like ANDREA, as well as advancing our understanding of the biological bases of complex behaviors and cognitions. Various other anomalous choice-related behaviors arising from specific disconnections and damage patterns are also described by Litt, Eliasmith and Thagard (2006). The enriched conceptions of decision phenomena provided by neural-level exploration allow for postulation of potential influences on behavior that would be missed by studies restricted to examining higher-level psychological processing.

A straightforward set of additional predictions deriving from our ANDREA simulations would be expectations of activation of the brain regions described in our model (Fig. 1) to be observed in imaging experiments involving people performing preference and choice tasks. But a more interesting explanatory and predictive utility of the model involves the study of *individual differences* in value functions. While the general features of the curve have been reproduced by many different experiments and in large numbers of people, the specific functions of different individuals are known to vary widely (e.g., Kahneman & Tversky, 1982). Because our simulations essentially represent the value function produced by a single brain (to which we have complete access, as its designers and builders) the model offers biological reasons for why and how individual value functions vary, while preserving some common general features. As discussed previously, connectivity strength between dorsal raphe serotonin and the amygdala can influence loss aversion, and thus the nature of the subjective valuations performed in orbitofrontal cortex. In persons with heightened serotonin sensitivity in the amygdala, we would expect to see a value function with an even steeper slope for losses than that of gains. The opposite also holds, in terms of lessened serotonergic influence or stronger midbrain dopaminergic connectivity, with either of these cases predictive of *less* discrepancy in slope between losses and gains (i.e., reduced loss aversion). The damaged innervation between the raphe and midbrain mentioned earlier would have the effect of steepening both the loss and gain portions of the curve, although the slope ratio might be unaffected. These connectivity manipulations can be understood psychologically as altering the extent to which one is emotionally aroused by losses and gains, with this affective influence central to producing subjective valuations of these deviations from a reference point.

We also predict that specific neural saturation range differences between brains may underlie individual differences in value function shape, with more readily saturating orbitofrontal populations prone to producing curves that level off quicker and more markedly. In the NEF, this saturation can itself be modeled asymmetrically around 0, so we can readily create a neural system of specific architecture and connectivity that yields significant leveling-off for losses but much less so for gains, or vice versa. In sum, the ANDREA model allows for numerous experimentally testable neural-level predictions regarding prospect theoretic behavior.

A NEURAL ACCOUNT OF DECISION AFFECT THEORY

The hedonic influence of prior expectations and counterfactual comparisons on the subjective valuation of outcomes is characterized by the work of Mellers and colleagues on what they call decision affect theory (Mellers & McGraw, 2001; Mellers, Schwartz, Ho, & Ritov, 1997; Mellers, Schwartz, & Ritov, 1999). Its fundamental claim is that evaluation by an individual of an outcome, event or decision option is strongly influenced by the "relative pleasure" it is considered to provide (Mellers, 2000). This relativity derives in part from the effects of counterfactual comparisons, as illustrated by the finding that Olympic silver medalists are more likely to feel disappointed than bronze medalists because of generally higher personal expectations (McGraw, Mellers, & Tetlock, 2005). Another factor is the degree to which an obtained outcome is considered surprising, with greater emotional impact for unexpected results (either good or bad) than for expected outcomes. The mathematical expression of decision affect theory is

$$R_O = J[u_O + d(u_O - u_E) * (1 - s_O)] \qquad [5]$$

(cf. equation [1] in McGraw, Mellers, & Tetlock, 2005). R_O is the emotional feeling associated with the obtained outcome and J is a linear function relating the felt pleasure to a specific

numerical response. u_O and u_E are the respective utilities of the obtained and expected outcomes, and $d(u_O - u_E)$ is a disappointment function that models how the obtained outcome is compared to the alternative expected outcome. s_O is the subjectively judged probability of the obtained outcome actually occurring, so the weighting by the complementary $(1 - s_O)$ term models the degree to which the obtained outcome was not expected (i.e., the subjective probability that something else would occur). The importance of emotional influence becomes clear with the finding that people will choose what *feels* best—that is, make decisions in such a manner as to maximize average positive emotional experience (Mellers, Schwartz, Ho, & Ritov, 1997).

Thus feelings about outcomes and choices, and hence the decisions people may be expected to make, are greatly influenced by the size and valence of the discrepancy between anticipated and actual results. The expected emotional reaction to gaining $20 will be vastly different if the prior expectation is gaining $100, versus the case where the expected yield is only $1. Indeed, the degree of influence of this discrepancy from anticipated results is such that an objectively worse outcome can sometimes be *more* pleasurable than one which is better. Consider the easily imaginable experience of feeling happier in stumbling across a twenty-dollar bill while walking home from work than from receiving an underwhelming one percent raise the same day, despite the monetary value of the raise being significantly higher than that of the found note. Decision affect theory thus describes in a revealing and systematic fashion the nature of certain situations in which less can actually feel like more. In describing a neural basis for the theory suggested by ANDREA, we shall draw upon an integration of the sort of framing discussed in our examination of prospect theory with the effects of the emotional context of information presentation on valuation and choice, which we now explore in depth.

Framing through Direct Emotional Arousal Influence

Our earlier discussion of framing effects focused on the type most discussed in prospect theory, concerning alternative descriptions of losses and gains. However, framing can affect decisions in other ways, as in the trolley-footbridge experiments of Greene and colleagues (2001). These experiments are not reference value manipulations, and are therefore not explainable by the neural mechanisms that we described for prospect theory. The judgment of morality reversal occurs between two outcomes that are *both* described as killing one person to save five others. The manipulation involved here is not one of reference point, but rather the personal or impersonal nature of the specific action performed that leads to the described outcome. It is thus differing contexts of choice presentation (i.e., the nature of the situation story) that produce the change in situational evaluation, rather than any suggestive presentation of choices in terms of either losses or gains. We will call this *emotional-context* framing, in contrast to the *reference-value* framing we discussed in relation to prospect theory.

We propose that emotional-context framing in the trolley-footbridge dilemma occurs through increased arousal associated with the direct, personalized action of pushing a person to their death, compared to the more detached and impersonal act of flipping a switch that will cause a trolley to divert from hitting five people towards hitting a single person. Such an increase in emotional engagement induces greater amplification of the subjective evaluation of causing a death in the personal case, which would provide a neurological basis for the reversal in typical judgments of the morality of the actions in question. fMRI experiments by Greene and colleagues seem to support this neural account of the trolley-footbridge dilemma, showing increased amygdala and orbitofrontal activity in cases of highly personal characterizations of morally debatable actions (Greene & Haidt, 2002; Greene et al., 2004).

Figure 8 illustrates the neural explanation we have described for the type of framing produced by changing the emotional context of the presented information. Just as for reference-value framing, we get different subjective valuations for scenarios that have identical objective values, which would allow for preference reversals to occur when the decision frame is altered. The primary difference is that this mechanism employs a *direct* manipulation of emotional arousal state, whereas our neural basis for the reference-value framing caused emotional modulation changes *indirectly* though manipulation of valuation-deviation reference points.

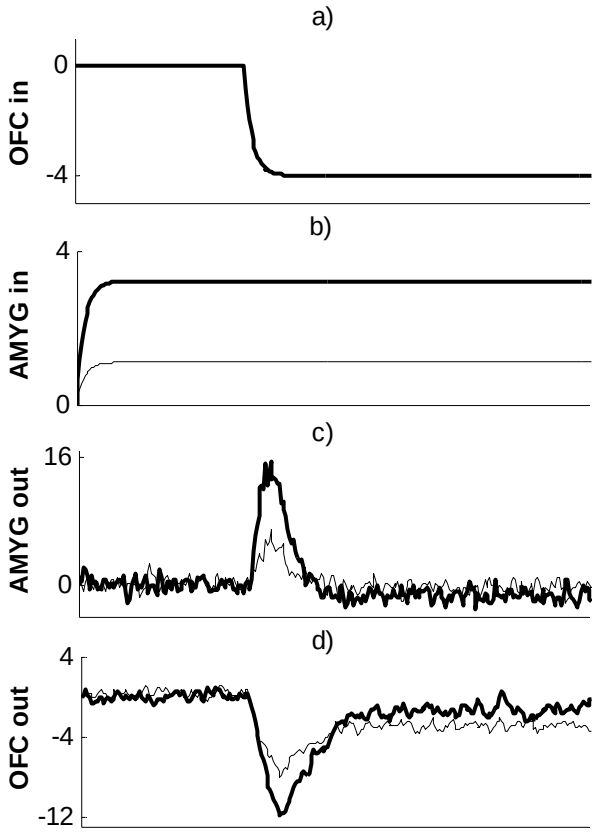

Figure 8: Simulation results for framing that changes emotional context. **a)** A step input to orbitofrontal cortex indicating a negative change in value, such as the consideration of a situation in which one's actions cause the death another person. **b)** Two different base arousal inputs to the amygdala, corresponding to differing levels of context-produced emotional engagement. **c)** Because of differing base arousal levels, arousal upsurges corresponding to the negative valuation deviation differ as well. **d)** Higher context-motivated base arousal leads to greater amplification of the subjective valuation change, and thus a belief that the more arousing scenario is actually worse than the objectively equivalent but less arousing scenario.

Decision Affect Theory as an Integrated Framing Phenomenon

Explanation of the experimental results examined in decision affect theory requires both reference-value framing and emotional-context framing. The hedonic impact of counterfactual comparisons can be produced by setting the reference point of the valuation deviation for an obtained outcome to the value of the unobtained counterfactual result, which is exactly reference-value framing. Then winning $20 when the counterfactual comparison is a $100 win would be construed as a loss, whereas a counterfactual comparison to winning only $1 would reframe the $20 win as a gain. In addition, the degree to which the obtained outcome was unexpected can have the effect of emotional-context framing. We formally model this in ANDREA by having emotional context influence neural behavior through the arousal input A_1 (see equation [4]) in a manner that takes into account the perceived probability of the obtained outcome X actually occurring:

$$A_1(t) = A_0(t) + \lambda \cdot (1 - \text{P[outcome } X\text{]}). \qquad [6]$$

In this enriched conception of emotional arousal, A_0 fills the previous role of A_1 as a base arousal level determined by external factors and provided as an input to the model. A_1 emotional arousal is now explicitly increased in inverse proportion to the expected probability of obtaining the outcome under consideration. Surprising, low-probability outcomes produce higher arousal than unsurprising outcomes, where $1 - P[X]$ is closer to 0. The constant multiplier λ may be related to the relative *affect-richness* of the outcome in question, as this variable seems strongly related to the degree to which uncertainty affects valuations (Rottenstreich & Hsee, 2001).

This direct manipulation of base emotional arousal is a formalization of emotional-context framing, in this case a context related to outcome uncertainty. It and reference-value framing, implemented in ANDREA by the mechanisms described earlier, together produce the

canonical finding of decision affect theory that objectively worse outcomes can sometimes feel better than more advantageous alternatives (Fig. 9). The mechanisms described above and illustrated in Figure 9 allow us to map the standard mathematical model of decision affect theory (equation [5]) onto specific neural structures and computations. The calculation of subjective utilities and their subsequent comparison, as embodied in the $d(u_O - u_E)$ term in equation [5], are performed neurologically through step-functional deviation in orbitofrontal cortex exactly as we modeled for prospect theoretic subjective valuation. The 'disappointment' effects of comparing obtained and expected outcome valuations result from prediction error computations by dopamine and serotonin networks feeding back to influence emotional arousal encoded in the amygdala, which in turn modulates orbitofrontal valuations. The subjective probability augmentation to our arousal representation (equation [6]) is identical to the $(1 - s_O)$ surprise term in the decision affect theory model of Mellers and colleagues. In combination with our multiplicative modulation of valuation by affective arousal (equation [1]), it is clear that this enhancement also has similar mathematical effects to that of for the surprise term in equation [5].

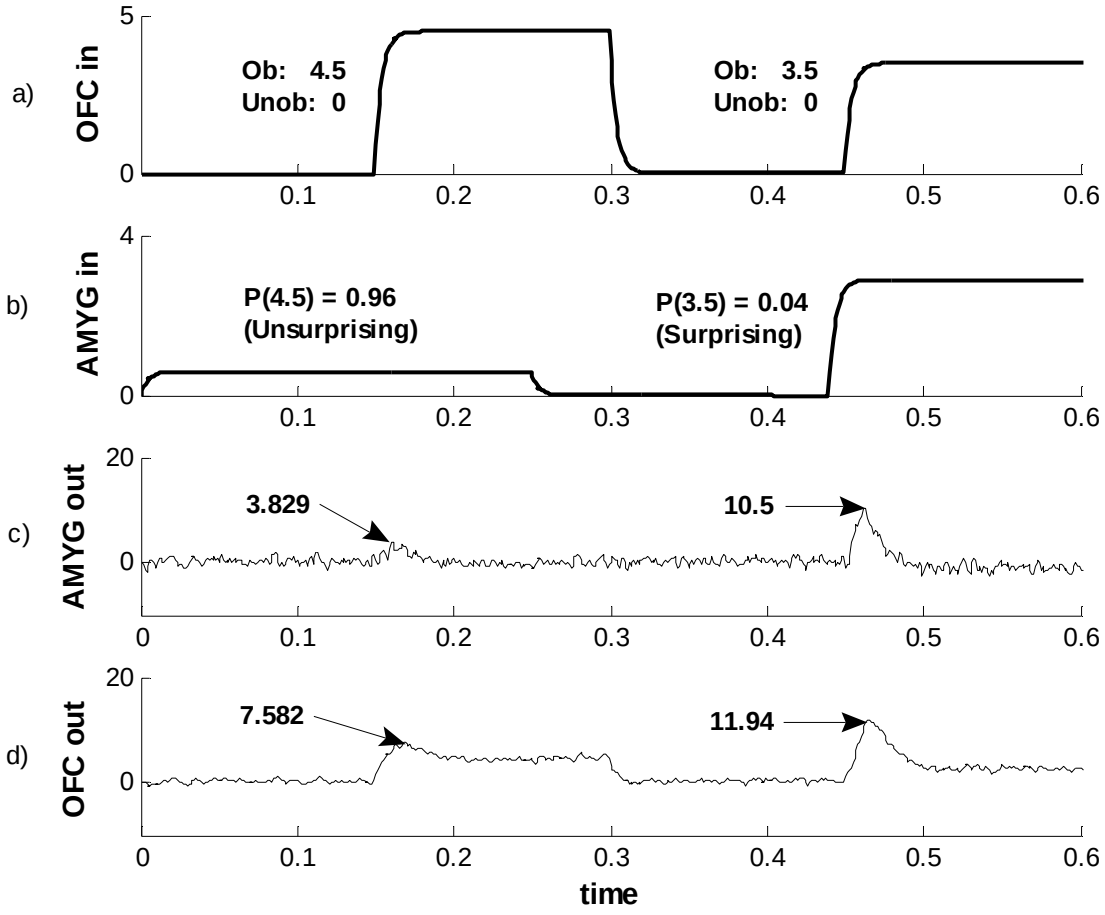

Figure 9: Decision affect theory as an integration of framing ideas. **a)** Counterfactual comparisons between obtained and unobtained outcomes are encoded through appropriate setting of valuation deviation reference points, as in reference-value framing. **b)** The surprising natures of obtained outcomes are encoded through direct manipulation of base emotional arousal input, as in emotional-context framing (equation [5]). **c)** The effect of surprise can outweigh an objectively larger valuation deviation, producing greater hedonic intensity for a surprising smaller gain than an unsurprising larger one in this case. **d)** The subjective valuation of an objectively worse outcome is greater than that of the better outcome, because of the added pleasure of being surprised by the smaller gain here.

Figures 10 and 11 show the results of more comprehensive simulations of the behavioral findings of Mellers and colleagues (Mellers, Schwartz, Ho, & Ritov, 1997). In Figure 10, the value of the unobtained counterfactual comparison outcome is held constant at $0 (i.e., neither losing nor gaining money) while the obtained outcome value and the expected probability of

obtaining that outcome are varied. Figure 11 describes an opposing experiment where the $0 is the unvarying *obtained* outcome—that is, subjects neither lose nor gain any money. What is instead varied is the expected probability of obtaining this $0 outcome and the value of the *unobtained* outcome used as a counterfactual comparison. In both figures and in both the behavioral and ANDREA simulation results, lower-probability curves (corresponding to surprising obtained outcomes) produce more intense affective experiences, as reflected by more extreme emotional response ratings. As well, there are cases in both the behavioral findings and our simulation data where an objectively worse outcome produces a more positive emotional response than one which is objectively greater. For instance, both Figures 10a and 10b show more elation from winning $17.50 instead of $0 with an expected probability of such a win of only 0.09 than for winning $31.50 instead of $0 when the anticipated probability of this outcome is 0.94. The surprising smaller gain feels better than the unexpected larger gain. Our proposed neural basis for decision affect theory thus provides a plausible and thorough biological characterization of the phenomenon.

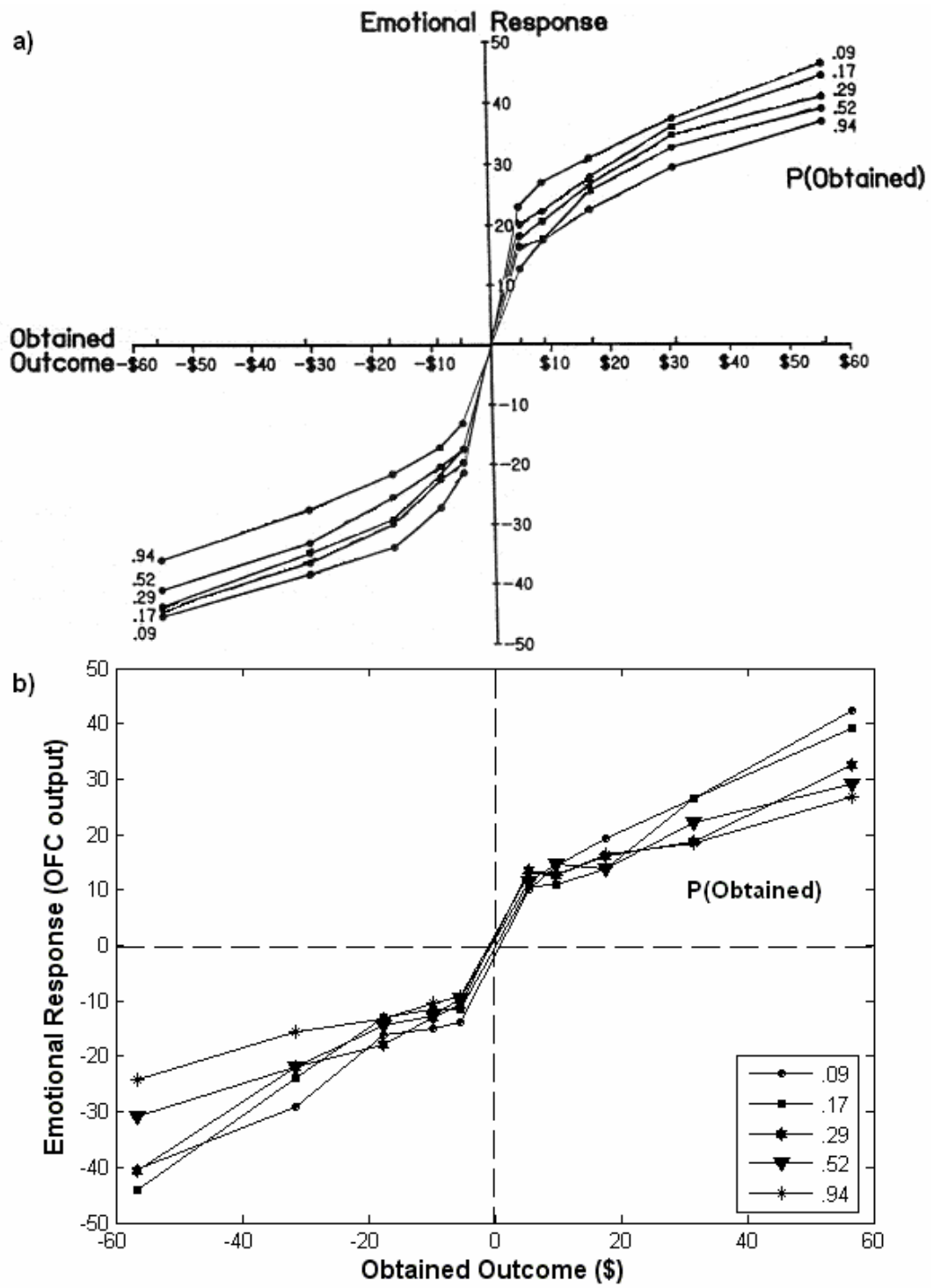

Figure 10: Comparing behavioral and simulation results in decision affect theory. **a)** Behavioral findings of Mellers et al., 1997, for lotteries with a constant $0 unobtained counterfactual comparison and varying *obtained* outcomes and expected obtained outcome probability. Emotional response was reported by subjects by ratings of feelings on a scale of 50 (extreme elation) to -50 (extreme disappointment) **b)** Results of model simulations of the Mellers et al. experiment in a), with data points determined through simulations in line with our proposed neural basis for decision affect theory (Fig. 9).

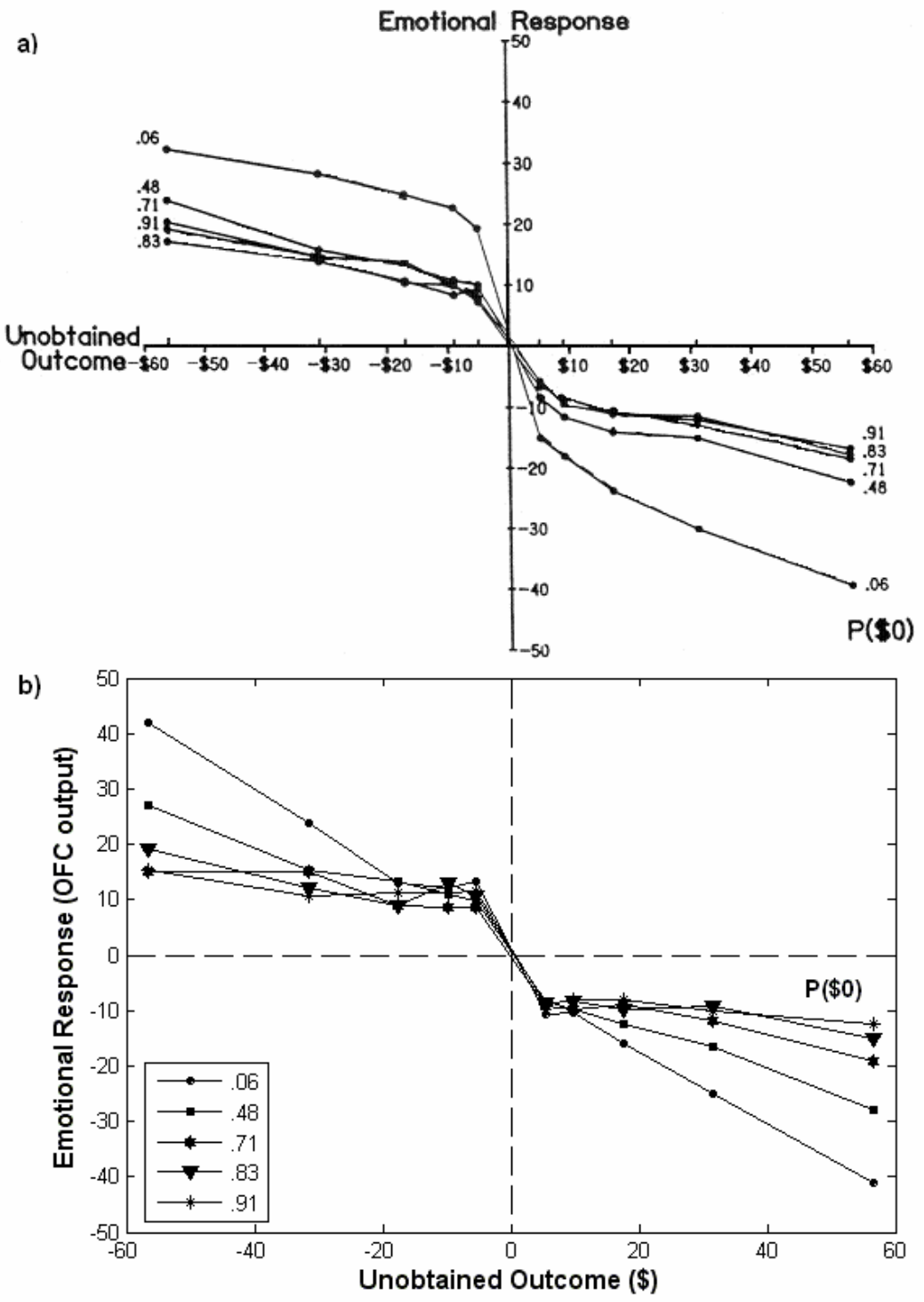

Figure 11: Comparing behavioral and simulation results in decision affect theory, opposite experiment to that described in Figure 10. **a)** Behavioral findings of Mellers et al., 1997, for lotteries with constant $0 obtained outcome and varying *unobtained* counterfactual comparison outcome value and expected obtained outcome probability. **b)** Model simulation of the experiment in a), with data points determined through simulations in line with our proposed neural basis for decision affect theory (Fig. 9).

We have been able to explain the central findings of decision affect theory using the same mechanisms that we applied to the phenomena explained by prospect theory, with the addition of framing by emotional context. A major motivating factor for the exploration of any subject at more basic levels of explanation is the desire to unite findings that are disconnected at higher levels of study through a set of shared lower-level mechanisms. In this vein, an important undertaking in the neuroscientific exploration of the psychology of preference and choice is to uncover shared underpinnings for phenomena that have yet to be rigorously tied together at the behavioral level. ANDREA demonstrates such a means to connect decision affect theory and prospect theory via the two neural mechanisms for framing we have outlined.

Predictions

There are undoubtedly other kinds of framing besides the reference-value and emotional-context types that we have discussed. Additional brain mechanisms may be required to explain other cases in which differing modes of identical information presentation produce divergent results, such as the case of reversals in preference when options are considered jointly versus separately (e.g., Hsee, Loewenstein, Blount, & Bazerman, 1999). Emotional-context framing could also be relevant to explaining a prominent result in constructive memory research. In the car accident study by Loftus and Palmer (1974), they describe significant effects on speed-of-impact memory judgments by subjects based simply on the emotiveness of the action verb used to describe the collision between two cars ("contacted each other" producing lower remembered speed judgments than "smashed into each other"). This seems analogous to the baseline emotional arousal manipulation inherent in our second neural framing mechanism, which in turn causes increasingly amplified subjective valuations that would correspond to inflated speed judgments by subjects when asked more emotively framed questions by Loftus and Palmer.

Further predictions arising from the framing mechanisms we have described relate to how the neural activity making up these mechanisms, and hence the effects of framing, could be induced without performing explicit decision framing. For instance, it might be possible to manipulate default subjective valuation reference points either upwards or downwards via positive or negative priming, or perhaps even through direct neuropharmacological intervention to influence orbitofrontal activity. This could potentially produce effects similar to reference-value framing in a less conspicuous manner. Similarly for emotional-context framing, manipulation of affective arousal in a manner wholly unrelated to situation context could cause "bleed-over" effects identical to those produced by situation-related arousal modulation. Methods of manipulation include prior exposure to violent or sexual imagery, relaxing or stressful preceding tasks, and direct pharmacological modulation of amygdala activity. There is a wide variety of means by which brain activity similar or identical to our mechanisms of framing can be induced either behaviorally or neurochemically, and we predict that such alternative routes should produce behaviors in people similar to explicit framing effects, regardless of whether or not they are aware of how they are being influenced.

GENERAL DISCUSSION

We have shown how neural affective decision theory, as stated in our four principles and as implemented in the ANDREA model, can account for the central phenomena described by prospect theory and decision affect theory. Our view of the general process of decision making is summarized in Figure 12. People are presented with a decision problem by verbal or perceptually experienced descriptions which they must interpret based on the context in which the decision is being made, resulting in an overall representation of the problem that is both cognitive and emotional. Options, outcomes, and goals can be encoded by verbal and other cognitive

representations, but with an ineliminable emotional content; in particular, goals are emotionally tagged. The translation of the presentation of a problem and its context into an internal cognitive-emotional representation produces framing effects, because different representations will invoke different neural-affective evaluations. ANDREA shows how these evaluations can be computed by coordinated activity among multiple brain areas, especially the orbitofrontal cortex, the amygdala, and dopamine and serotonin systems involved in encoding positive and negative changes in valuation. The result is decisions that select options inducing the highest emotional subjective valuations.

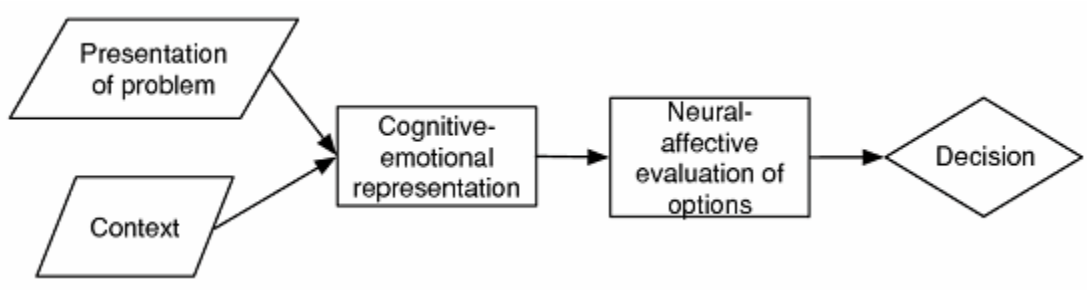

Figure 12. Overview of neural affective decision theory.

ANDREA has greater explanatory scope than other neurocomputational models of decision and reward that have focused on in-depth modeling of more restricted subsystems of the brain, and accordingly limited ranges of behavioral phenomena. Two such examples are the model of reward association reversal in orbitofrontal cortex by Deco and Rolls (2005) and the GAGE model of cognitive-affective integration in the Iowa gambling task and self-evaluations of physiologically ambiguous emotional states (Wagar & Thagard, 2004). Additionally, a straightforward diffusion decision process implemented in superior colliculus cells seems able to accurately characterize accuracy and reaction time in simple two-choice decision tasks (Ratcliff, Cherian, & Segraves, 2003). Task modeling of this sort is important for exploring basic details of neural mechanisms for specific phenomena, but examining brain processes on a larger scale is

required for explaining more complex and wide-ranging psychological findings, such as prospect theory and decision affect theory. Busemeyer and Johnson (2004) describe a connectionist model that they apply to a range of behaviors as diverse as those explored by ANDREA, including preference reversal effects and loss aversion. The network model called affective balance theory (Grossberg & Gutowski, 1987) also explores a wide range of risky decision phenomena in a mathematically sophisticated fashion, and proposes effects of emotional context on cognitive processing that are largely consistent with those implemented in ANDREA. The main improvement that our approach offers over these two models is in neurological realism, as reflected by modeled characteristics of individual processing units ('neurons') and the mapping of proposed computations onto specific brain regions and interactions supported by empirical findings. The models of Grossberg and Gutowski (1987) and Busemeyer and Johnson (2004) are not comparable in this respect to either ANDREA or the previously mentioned works of Deco and Rolls (2005) and Wagar and Thagard (2004). This of course is not a criticism of such methods of modeling. Rather, it is more of an indication of the different levels at which theoreticians can formulate explanations of behaviorally studied psychological phenomena. While the proposed mechanisms are interesting from the larger perspective of computational models of decision making, these artificial neural networks are of a fundamentally different nature than ANDREA and similar models in computational cognitive neuroscience.

A recent model of decision-related interactions between basal ganglia and orbitofrontal cortex by Frank and Claus (2006) recognizes the utility of taking the sort of broad-scale approach we employ in our design of ANDREA, and maintains a similar level of neurological realism and detail. Despite certain similarities regarding modeled brain regions and proposed computations, this model diverges from ANDREA in several fundamental ways, leading to both

different mechanisms and different explanatory targets for each of our models. After briefly describing the structure of the Frank and Claus (2006) model, we compare it in detail with our own model of valuation and choice phenomena.

The main focus of the Frank and Claus model is the means by which basal ganglia dopaminergic activity and orbitofrontal computations enable adaptive decision making responsive to contextual information. Computation and representation of expected decision value information is accomplished through a division of labor between subcortical dopamine and prefrontal networks. A basal ganglia dopaminergic network learns to make decisions based on the relative probability of such decisions leading to positive outcomes. This process is augmented by orbitofrontal circuits that provide a working memory representation of associated reinforcer value magnitudes that exercises top-down control on the basal ganglia activity, which allows more flexible response to rapidly changing inputs. The proposed computations are detailed, elegant and well-supported by empirical data, and the model is effective in explaining decision-related behaviors as diverse as risk aversion/seeking, reversal learning, and peoples' performance in a variant of the Iowa gambling task in both normal and brain-damaged scenarios.

While ANDREA does not implement the specific computations proposed by Frank and Claus (2006), this is due more to differing targets of explanation than to any major inconsistencies in our respective conceptions of the roles of various brain regions in decision making. Both models describe important functional differences between orbitofrontal-amygdala networks and dopaminergic activity in line with empirical findings demonstrating the involvement of orbitofrontal cortex in valuation and dopaminergic encoding of reward prediction error. How these subsystems might interact is a question that both ANDREA and Frank and Claus (2006) address, and one that has been neglected in previous theoretical modeling of the

neurobiology of reward. Nevertheless, whereas Frank and Claus develop a comprehensive characterization of how orbitofrontal cortex learns and represents reinforcer value, our goal is to describe how specific external influences *differentially alter* the magnitudes of these orbitofrontal-encoded values, such as via the asymmetric emotional modulation by losses and gains on valuations that we describe as the basis of loss aversion in prospect theory. As a result, the models are best suited to providing neural explanations of different psychological phenomena, and where they address similar phenomena they do so with contrasting emphases on specific relevant brain mechanisms.

The most prominent difference is that of representational complexity of amygdala activity. In both models, the amygdala encodes the magnitude of losses and gains in proportion to overall activity level, which then influences orbitofrontal representations of reward values. Frank and Claus (2006) do not explore how the amygdala forms such representations of reinforcer magnitude, providing them instead as direct model inputs. In contrast, ANDREA models multifaceted means by which emotional arousal related to outcome magnitude is encoded by the amygdala (equations [4] and [5]). This allows for the postulation of neural explanations for phenomena not addressed by Frank and Claus (2006), such as multiple mechanisms for framing and the observations of decision affect theory. In addition, while both models describe loss aversion as resulting from greater amygdala activation by losses than equivalent gains, ANDREA offers specific neurological reasons of how this might occur through differential calibration of distinct loss and gain reward prediction error networks, as well as feedback to the amygdala from dorsolateral and cingulate processing of behavioral saliency. Our detailed characterization of how the amygdala comes to represent magnitude information thus allows us

to both explain additional phenomena and provide more complete biological mechanisms for valuation and decision behaviors simulated by both models.

Important structural differences between the models result from their differing conceptions of amygdala activity. The Frank and Claus (2006) model focuses primarily on the top-down biasing effect of orbitofrontal activity on gradual, multi-trial learning in the basal ganglia dopamine network, with much less emphasis on the reciprocal influence of relative reinforcement probability computations on orbitofrontal activation. We are able to explore in depth such effects in ANDREA through feedback of reward prediction error information to the amygdala that influences its encoding of emotional arousal, which in turn modulates orbitofrontal valuations. These valuations have top-down biasing effects on reinforcement learning and action selection similar to those modeled by Frank and Claus (2006), for instance how differing orbitofrontal representations of identical reward values caused by framing effects produce different activity in dopamine and serotonin prediction error subsystems.

There are several other issues on which the models differ in important ways. Perhaps the most obvious is regarding the role of serotonin acting in opponency with dopamine in reward prediction error. Frank and Claus (2006) argue that the low baseline firing rates of dopaminergic cells need not mean that firing rate depressions are less capable of encoding highly negative outcomes, as there may exist countervailing sensitivity differences in dopamine receptors to firing bursts versus dips. While they remain open to a role for serotonin in negative reinforcement learning and aversive stimulus processing mediated within OFC, they argue for the centrality of dopamine firing dips to negative valuation processing mediated by the basal ganglia, even in light of evidence showing these dips are physiologically limited in scope (e.g., Bayer & Glimcher, 2005). Our model can be considered structurally consistent with this line of

reasoning, since the mutual biasing effects we have implemented between dopamine and serotonin produce the same dopaminergic firing depressions utilized computationally by Frank and Claus in cases where we describe encoding via serotonergic activity. Clearly, though, the models differ markedly in the degree to which they assign explanatory import to dopamine firing dips versus concomitant serotonergic firing increases. Further differences are evident in respective extents of brain region modeling. Frank and Claus employ more complex conceptions of orbitofrontal cortex and midbrain dopamine areas than are implemented by ANDREA, differentially utilizing specific subpopulations of these broadly defined brain areas. In contrast, ANDREA includes limited but important contributions from anterior cingulate and dorsolateral prefrontal cortices. These include involvement in encoding the behavioral relevance of outcomes, how this encoding may differ for positive and negative outcomes, and the subsequent effects of behavioral saliency on emotional arousal. These are brain regions omitted in the model of Frank and Claus (2006) that they acknowledge may be crucial to understanding decision phenomena. Finally, ANDREA is the first model to explore a possible role for neural saturation in explaining the nature of subjective valuation. This effect is not examined in the Frank and Claus work, which does not address the leveling off of the prospect theory value function for increasingly extreme losses and gains. Thus, while these two models are fairly consistent with one another and share similarities in their large-scale approaches to modeling the neural foundations of decision making, they both make unique contributions to explaining different aspects of relevant behaviors and psychological processes.

One of the most fertile areas for future applications of neural affective decision theory and the ANDREA model is the burgeoning field of neuroeconomics, which operates at the intersection of economics, psychology, and neuroscience (Camerer, Loewenstein, & Prelec,

2005; Glimcher & Rustichini, 2004; Sanfey, Loewenstein, McClure, & Cohen, 2006). Examples of such applications include the previously mentioned findings regarding preference reversal in joint versus separate option evaluation (Hsee, Loewenstein, Blount, & Bazerman, 1999) and observed interactions between risk, uncertainty and emotion (Rottenstreich & Hsee, 2001), both of which seem explainable via the neurological mechanisms we have modeled. Unlike traditional economic theory, we do not take preferences as given, but rather explain them as the result of specific neural operations. A person's preference for *A* over *B* is the result of a neural-affective evaluation in which the representation of *A* produces a more positive anticipated reward value (or at least a less negative value) than the representation of *B*. As depicted in Figure 12, the neural-affective evaluation of options depends on their cognitive-emotional representation, which can vary depending on the presentation and context of information. This dependence explains why actual human preferences often do not obey the axioms of traditional microeconomic theory. In addition to neuroeconomics, we are exploring the relevance of our theory and model to understanding ethical judgments, the neural bases of which are under increasing investigation (Casebeer and Churchland, 2003; Greene and Haidt, 2002; Moll et al., 2005). Finally, while neural affective decision theory is primarily intended as a descriptive account of how people actually do make decisions, but it can provide a starting point for developing a prescriptive theory of how they ought to make better decisions (Thagard, 2006).

Like all models, ANDREA provides a drastically simplified picture of the phenomena it simulates, and there are many possible areas for improvement and extension. These include increasing the complexity of individual populations, adding more brain areas, modeling more relationships between brain areas, and exploring the effects of neuronal firing saturation beyond simply orbitofrontal cortex. Nevertheless, we have described a variety of neurobiologically

realistic mechanisms for fundamental decision processes, and shown their applicability to explaining several major experimental findings in behavioral decision research. Besides implementing original mechanistic ideas, such as a role for saturation in explaining diminishing marginal sensitivity in prospect theory, ANDREA contributes to two important classes of explanation in decision neuroscience: 1) *Generalization* and novel *syntheses* of hitherto unrelated mechanisms of neural processing (e.g., multiplicative models of attention and reward reinforcement learning); and 2) Specific and detailed *grounding* of behaviorally explored psychological phenomena in such plausible and realistic neurocomputational mechanisms. The result, we hope, is a deeper understanding of how and why people make the choices that they do.

ACKNOWLEDGMENTS

We thank Daniela O'Neill, Christopher Parisien, Bryan Tripp, and three anonymous reviewers for comments on earlier versions. This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada.

APPENDIX A: NEUROCOMPUTATIONAL DETAILS

Our reward model was implemented in MATLAB 7.0.1 on a PC with an Intel Pentium 4 processor running at 2.53 GHz, with 1.00 GB of RAM available. For simulations of the extent that we have conducted and described here, these specifications represent close to the minimum required, based on the memory and other resource requirements of the most recent version of the *NESim* NEF simulation software running within MATLAB. *NESim* documentation and software download links are available online at http://compneuro.uwaterloo.ca.

We modeled spiking activity for a total of 7600 neurons spread over 7 specific populations (Fig. 1), using the common and physiologically realistic leaky integrate-and-fire

(LIF) model for each of our modeled neurons (see Appendix B). In particular, we use 800-1200 neurons for simulating each of the amygdala, orbitofrontal cortex, ventral striatum, anterior cingulate cortex, and dorsolateral prefrontal cortex, representing one- to three-dimensional vectors as needed in the neural engineering framework (Appendix B). The areas representing midbrain dopaminergic neurons and the dorsal raphe nucleus of the brainstem are each modeled with 1200-neuron ensembles, each with several discrete subpopulations, in order to capture the additional complexities involved in the encodings and transformations we assign to these areas in our model (recurrent, rectified, biased-opponent calculation of reward prediction error; see Appendix C).

Each individual neuron is based on a reduced-complexity biophysical model that includes features fundamental to most neurons, including conventional action potentials (spikes), spike train variability, absolute refractoriness, background firing, and membrane leak currents. The membrane time constant for our LIF neurons is set to 10 ms, with a refractory period (a post-spike delay during which the neuron may not fire) of 1 ms. We introduce at simulation runtime 10% Gaussian, independent zero-mean noise, relative to normalized maximum firing rates, to simulate the noisy environment in which our neurons operate. These choices are based on plausible biological assumptions (see Eliasmith & Anderson, 2003), and we have made reasonable efforts to select neurobiologically realistic values for other network cell parameters as well. For example, our 5 ms synaptic time constant for dorsal raphe serotonergic neurons is consistent with the 3 ms decay constants observed *in vivo* (Li & Bayliss, 1998). Neurons in the cortical areas we have modeled have been shown to have maximum firing rates ranging from 20-40 Hz in dorsolateral prefrontal and orbitofrontal areas (Wallis & Miller, 2003) to at least 50 Hz in anterior cingulate cortex (Davis et al., 2000). Thus, our selection of a 10-80 Hz saturation

range for neurons in these model ensembles is a physiologically reasonable compromise to maintain population sizes that are manageable for simulation purposes. Ensembles must grow increasingly large to allow for meaningful representation with slower-firing, smaller saturation-range neurons making up the ensembles.

Practical considerations nonetheless made necessary some limitations on biological realism. Principally, the saturation ranges we selected for our modeled subcortical regions are appreciably higher (by roughly a factor of 10) than those observed empirically for typical neurons in these areas. The much larger neural ensemble sizes that would be required for clean representation and transformation using the extremely low experimentally observed firing rates would have made our large-scale simulations computationally impracticable given available resources. We additionally support this compromise of biological realism by noting that significantly higher firing peaks (greater than 100 Hz) are observed in the *bursting* behavior of both certain raphe serotonergic neurons (e.g., Gartside et al., 2000; Hajós et al., 1995) and subpopulations of the amygdala (Driesang & Pape, 2000; Paré & Gaudreau, 1996), and to less degree in midbrain and striatum dopaminergic neurons as well (Hyland et al., 2002). The specific activity our model produces in these subcortical areas seems well-suited to coding via bursting neurons (large but transient firing upsurges that interpose lengthier periods of near-zero activity). While we do not define here an explicit alternative neuron model to LIF, Eliasmith (2005a) describes how bursting could be incorporated into the NEF, and thus *NESim* simulations.

For the most part, we have chosen neuron firing thresholds (that is, the respective input levels above which individual neurons begin to respond) from an even distribution over a range represented symmetrically around zero, with neuron preferred directions in the space of representation also chosen from an even distribution (i.e., equal numbers of ‘on’ and ‘off’

neurons; see Appendix B). The exceptions to this rule are single subpopulations of both the dorsal raphe and midbrain dopaminergic regions. In these cases, our establishment of minimum firing thresholds of zero combines with an exclusive use of positively sloped 'on' neurons in these subpopulations to produce insensitivity to negative values (i.e., rectified reward prediction error encoding). This corresponds well with experimentally observed physiological limitations in the computation of reward prediction error in these brain regions (see Bayer & Glimcher, 2005, as well as the discussion of dopamine and serotonin within the main text description of the ANDREA model).

APPENDIX B: NEF REPRESENTATION, TRANSFORMATION AND DYNAMICS

The NEF consists primarily of three fundamental principles regarding the representation of information in neural populations, the means by which these representations are transformed through interactions between populations, and the control theoretic nature of the characterization of neural dynamics. Eliasmith and Anderson (2003) present a rigorous explication and analysis of the NEF, but the mathematical details we outline here should be sufficient for understanding the fundamental nature and operation of our neural model of affective choice and valuation.

Consider a neural ensemble whose activities $a_i(\mathbf{x})$ *encode* some vector quantity $\mathbf{x}(t)$ mapping onto a real-world quantity (eye position, emotional arousal, etc.). Note that this quantity need not be a vector; scalars, functions, and function spaces can be represented and manipulated in the NEF in a near-identical fashion. The encoding of $\mathbf{x}$ involves a conversion of $\mathbf{x}(t)$ into a neural spike train:

$$a_i(\mathbf{x}) = \sum_n \delta(t - t_{in}) = G_i\left[J_i(\mathbf{x}(t))\right], \qquad \text{[B1]}$$

where $G_i[\cdot]$ is the nonlinear function describing the specific nature of the spiking response, J_i is the current in the cell body (soma) of a particular neuron and *i* and *n* are relevant indices (*i*

indexing specific neurons, *n* indexing the individual spikes produced by a given neuron). The nonlinearity *G* we employ is the common leaky-integrate-and-fire (LIF) model:

$$dV_i/dt = -\left(V_i - J_i(\mathbf{x})R\right)/\tau_i^{RC}, \qquad \text{[B2]}$$

where V_i represents somatic voltage, R the leak resistance, and τ_i^{RC} the RC (membrane) time constant. The system is integrated until the membrane potential V_i crosses the threshold V_{th}, at which point a spike $\delta(t-t_{in})$ is generated and V_i is reset to zero for the duration of the refractory period, τ_i^{ref} (Eliasmith & Anderson, 2003). A basic description of the soma current is

$$J_i(\mathbf{x}) = \alpha_i \left\langle \tilde{\phi}_i \cdot \mathbf{x} \right\rangle + J_i^{bias} + \eta_i, \qquad \text{[B3]}$$

where $J_i(\mathbf{x})$ is the current input to neuron *i*, **x** is (in this case) the vector variable of the stimulus space encoded by the neuron, α_i is a gain factor, $\tilde{\phi}_i$ is the preferred direction vector of the neuron in the stimulus space, J_i^{bias} is a bias current (accounting for any background activity) and η_i models any noise to which the system is subject. Note in particular that the *dot product* $\left\langle \tilde{\phi}_i \cdot \mathbf{x} \right\rangle$ describes how a potentially complex (i.e., high-dimensional) physical quantity, such as an encoded stimulus, is related to a scalar signal describing the input current. For scalars, the encoding vector is either +1 (an 'on' neuron) or -1 (an 'off' neuron). [B1] thus captures the nonlinear encoding process from a high-dimensional variable, **x**, to a one dimensional soma current, J_i, to a train of neural spikes, $\delta(t-t_{in})$.

Under this encoding paradigm, the original stimulus vector representation can be estimated by *decoding* those activities; that is, converting neural spike trains back into quantities relevant for explanations of neural computation at the level of our chosen representations. A plausible means of characterizing this decoding is as a specific *linear* transformation of the spike

train. In the NEF, the original stimulus vector $\mathbf{x}(t)$ is decoded by computing an estimate $\hat{\mathbf{x}}(t)$ using a linear combination of filters $h_i(t)$ that are weighted by certain decoding weights ϕ_i :

$$\hat{\mathbf{x}}(t) = \sum_{in} \delta(t - t_{in}) * h_i(t)\phi_i = \sum_{in} h_i(t - t_{in})\phi_i \ , \qquad [B4]$$

where the decoding weights are calculated by a mean-squared error minimization (Eliasmith & Anderson, 2003) and the operation '$*$' indicates convolution. The $h_i(t)$ filters are linear temporal decoders, which are taken to be the postsynaptic currents (PSCs) in the associated neuron *i* for reasons of biological plausibility. Together, the nonlinear encoding in [B1] and the linear decoding in [B4] define an ensemble 'code' for the neural representation of **x**.

The next aspect of the NEF to examine is the means by which computations are performed in order to transform the representations present in a given model. The main task needed to be performed is the calculation of connection weights between the different populations involved in a transformation. As an example, let us consider the transformation $z = x \cdot y$. The process of connection weight calculation can be characterized as substituting into our *encoding* equation [B1] the *decodings* of *x* and *y* (as per [B4]) in order to find the encoding of *z*, which represents our transformation of interest:

$$\begin{aligned} c_k(z) &= c_k(x \cdot y) \\ &= G_k\left[\alpha_k \tilde{\phi}_k (x \cdot y) + J_k^{bias}\right] \\ &= G_k\left[\alpha_k \left(\tilde{\phi}_k \sum_i a_i(x)\phi_i^x \cdot \sum_j b_j(y)\phi_j^y\right) + J_k^{bias}\right] \\ &= G_k\left[\sum_{i,j} \omega_{kij} a_i(x) b_j(y) + J_k^{bias}\right], \end{aligned}$$

where $\omega_{kij} = \alpha_k \tilde{\phi}_k \phi_i^x \phi_j^y$ represents the connection weights between neurons *i*, *j* and *k* in the *x*, *y*, and *z* populations, respectively. It should be noted that the nonlinear neural activity interaction

suggested in this example is avoided in our actual model—all interactions are implemented in a purely linear fashion, as is typically taken to be the case in real neural systems (see Eliasmith & Anderson, 2003, for complete implementation details).

Finally, dynamics play a fundamental role in the overall operation of our model, such as in our recurrent reward prediction error computation. We can describe the dynamics of a neural population in control theoretic form via the dynamics state equation that is at the basis of modern control theory:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \text{[B5]}$$

where **A** is the dynamics matrix, **B** is the input matrix, **u**(*t*) is the input or control vector, and **x**(*t*) is the state vector. At this high-level of characterization we are detached from any neural-level implementation details. It is possible, however, to introduce simple modifications that render the system neurally plausible. The first step in converting this characterization is to account for intrinsic neural dynamics. To do so, we assume a standard PSC model given by $h(t) = \tau^{-1} e^{-t/\tau}$, and then employ the following derived relation (Eliasmith & Anderson, 2003):

$$\begin{aligned} \mathbf{A}' &= \tau\mathbf{A} + \mathbf{I} \\ \mathbf{B}' &= \tau\mathbf{B} \end{aligned} \qquad \text{[B6]}$$

so that our *neurally plausible* high-level dynamics characterization becomes

$$\mathbf{x}(t) = h(t) * \left[\mathbf{A}'\mathbf{x}(t) + \mathbf{B}'\mathbf{u}(t)\right]. \qquad \text{[B7]}$$

To integrate this dynamics description with the neural representation code we described previously, we combine the dynamics of [B7], the encoding of [B1], and the population decoding of **x** and **u** as per [B4]. That is, we take decodings $\hat{\mathbf{x}} = \sum_{jn} h_j(t - t_{jn})\phi_j^{\mathbf{x}}$ and $\hat{\mathbf{u}} = \sum_{kn} h_k(t - t_{kn})\phi_k^{\mathbf{u}}$ and introduce neural dynamics into the encoding operation as follows:

$$\begin{aligned}\sum_n \delta(t-t_{in}) &= G_i\left[\alpha_i\left\langle\tilde{\phi}_i\mathbf{x}(t)\right\rangle + J_i^{bias}\right] \\ &= G_i\left[\alpha_i\left\langle\tilde{\phi}_i\left[\mathbf{A}'\hat{\mathbf{x}}(t)+\mathbf{B}'\hat{\mathbf{u}}(t)\right]\right\rangle + J_i^{bias}\right] \\ &= G_i\left[\alpha_i\left\langle\tilde{\phi}_i\left[\mathbf{A}'\sum\nolimits_{jn} h_j(t-t_{jn})\phi_j^{\mathbf{x}} + \mathbf{B}'\sum\nolimits_{kn} h_k(t-t_{kn})\phi_k^{\mathbf{u}}\right]\right\rangle + J_i^{bias}\right] \\ &= G_i\left[\sum\nolimits_{jn}\omega_{ij}h_j(t-t_{jn}) + \sum\nolimits_{kn}\omega_{ik}h_k(t-t_{kn}) + J_i^{bias}\right].\end{aligned} \qquad \text{[B8]}$$

It is interesting to note that $h(t)$ in the above characterization defines both the neural dynamics *and* the decoding of the relevant representations. $\omega_{ij} = \alpha_i\left\langle\tilde{\phi}_i\mathbf{A}'\phi_j^{\mathbf{x}}\right\rangle$ and $\omega_{ik} = \alpha_i\left\langle\tilde{\phi}_i\mathbf{B}'\phi_k^{\mathbf{u}}\right\rangle$ describe the recurrent and input connection weights, respectively, which implement the dynamics defined by the control theoretic structure from [B7] in a neurally plausible network.

APPENDIX C: REPRESENTATION/TRANSFORMATION SUMMARY

Table C1 encapsulates the complete inputs, outputs and transformations we use to model specific interactions between the brain regions included in our model (see also Figure 1 and discussions of equations in the main body for more high-level, conceptual characterizations of connectivity and signal transformation). Variable names are as in the text of the Methods section.

Brain area	**Inputs**	**Outputs**
AMYG	$A_0(t)$ (ext.) $A_1(t)$ $DA(t)$ $5\text{-}HT(t)$ $C(t)$	$A(t) = A_1(t) + \beta \cdot DA(t) + \gamma \cdot 5\text{-}HT(t) + C(t)$, where $A_1(t) = A_0(t) + \lambda \cdot (1 - \text{P[outcome } X\text{]})$
OFC	$V(t)$ (ext.) $A(t)$	$S(t) = A(t) \cdot V(t)$
5-HT	$S(t)$ $E^+(t)$ $P(t-1)$	$E^-(t) = P(t-1) - S(t)$ $5\text{-}HT(t) = \sigma E^-(t) - (1-\sigma)E^+(t)$ $P(t) = P(t-1) + \alpha E(t)$
DA	$S(t)$ $E^-(t)$ $P(t-1)$	$E^+(t) = S(t) - P(t-1)$ $DA(t) = \sigma E^+(t) - (1-\sigma)E^-(t)$ $P(t) = P(t-1) + \alpha E(t)$
VS	$DA(t)$ $5\text{-}HT(t)$	$E(t) = DA(t) - 5\text{-}HT(t)$
ACC	$S(t)$ $E(t)$ $C(t)$	$B(t) = 2 \cdot (S(t) \geq 0) - 1$ $R(t) = B(t) / [\eta + E(t)]$ $C(t)$
DLPFC	$R(t)$ $5\text{-}HT(t)$	$C(t) = \mu \cdot 5\text{-}HT(t)$

Table C1: Transformation summary. A_0 and V are provided as external inputs to the model. Note the recurrent connectivity and opponent interaction between 5-HT and DA. Abbreviations: AMYG, amygdala; OFC, orbitofrontal cortex; 5-HT, raphe dorsalis serotonergic neurons; DA, midbrain dopaminergic neurons; VS, ventral striatum; ACC, anterior cingulate cortex; DLPFC, dorsolateral prefrontal cortex.

These equations describe explicitly the nature of the connectivity relationships and signal transformation processes outlined in Figure 1 and discussed in the main body description of the ANDREA model.

REFERENCES

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature,* 433, 68-72.

Bayer, H. M. & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47, 129-141.

Bechara, A., Damasio, H., & Damasio, A. R. (2003). Role of the amygdala in decision-making. *Annals of the New York Academy of Sciences, 985*, 356-369.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, 10(3), 295-307.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A. & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, 30, 619-639.

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 133-154). Oxford: Blackwell.

Bush, G., Luu, P. & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6), 215-222.

Camerer, C. F. (2000). Prospect theory in the wild: evidence from the field. In D. Kahneman & A. Tversky (Eds.), *Choices, Values, and Frames*. New York: Cambridge University Press.

Camerer, C., Loewenstein, G. F., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature, 34*, 9-64.

Casebeer, W. D., & Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach. *Biology and philosophy, 18*, 169-194.

Churchland, P. S. (1996). Feeling reasons. In A. R. Damasio, H. Damasio & Y. Christen (Eds.), *Neurobiology of Decision-Making* (pp. 181-199). Berlin: Springer.

Cools, R., Calder, A. J., Lawrence, A. D., Clark, L., Bullmore, E., & Robbins, T. W. (2005). Individual differences in threat sensitivity predict serotonergic modulation of amygdala response to fearful faces. *Psychopharmacology*, 180(4), 670-679.

Damasio, A. R. (1994). *Descartes' error*. New York: G. P. Putnam's Sons.

Davis, K. D., Hutchison, W. D., Lozano, A. M., Tasker, R. R., & Dostrovsky, J. O. (2000). Human anterior cingulate cortex neurons modulated by attention-demanding tasks. *Journal of Neurophysiology*, 83(6), 3575-3577.

Davis, K. D., Taylor, K. S., Hutchison, W. D., Dostrovsky, J. O., McAndrews, M. P., & Richter, E. O. et al. (2005). Human anterior cingulate cortex neurons encode cognitive and emotional demands. *Journal of Neuroscience, 25*(37), 8402-8406.

Daw, N. D., Kakade, S. & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, 15, 603-616.

Deakin, J. F. W. (1983). Roles of brain serotonergic neurons in escape, avoidance and other behaviors. *Journal of Psychopharmacology*, 43, 563-577.

Deco, G. & Rolls, E. T. (2005). Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cerebral Cortex*, 15, 15-30.

Driesang, R. B. & Pape, H. (2000). Spike doublets in neurons of the lateral amygdala: mechanisms and contribution to rhythmic activity. *NeuroReport,* 11(8), 1703-1708.

Eliasmith, C. (2005a). A unified approach to building and controlling spiking attractor networks. *Neural Computation*, 17(6), 1276-1314.

Eliasmith, C. (2005b). Cognition with neurons: a large-scale, biologically realistic model of the Wason task. In B. Bara, L. Barasalou & M. Bucciarelli (Eds.), *Proceedings of the XXVII Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eliasmith, C. & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C. & Anderson, C. H. (2000). Rethinking central pattern generators: a general framework. *Neurocomputing,* 32-33(1-4), 735-740.

Evenden, J. L. & Ryan, C. N. (1996). The pharmacology of impulsive behaviour in rats: the effects of drugs on response choice with varying delays of reinforcement. *Psychopharmacology (Berl.)*, 128, 161-170.

Gartside, S. E., Hajos-Korcsok, E., Bagdy, E., Harsing, L. G., Sharp, T., & Hajós, M. (2000). Neurochemical and electrophysiological studies on the functional significance of burst firing in serotonergic neurons. *Neuroscience,* 98(2), 295-300.

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

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science,* 293, 2105-2108.

Greene, J. & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517-523.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44, 389-400.

Grossberg, S. & Gutowski, W. E. (1987). Neural dynamics of decision making under risk: affective balance and cognitive-emotional interactions. *Psychological Review*, 94(3), 300-318.

Hajós, M., Gartside, S. E., Villa, A. E., & Sharp, T. (1995). Evidence for a repetitive (burst) firing pattern in a sub-population of 5-hydroxytryptamine neurons in the dorsal and median raphe nuclei of the rat. *Neuroscience*, 69(1), 189-197.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96, 651-656.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: a review and theoretical analysis. *Psychological Bulletin*, 125(5), 576-590.

Hsee, C. K., Rottenstreich, Y., & Xiao, Z. (2005). When is more better? On the relationship between magnitude and subjective value. *Current Directions in Psychological Science*, 14, 234-237.

Hyland, B. I., Reynolds, J. N., Hay, J., Perk, C. G., & Miller, R. (2002). Firing modes of midbrain dopamine cells in the freely moving rat. *Neuroscience*, 114(2), 475-492.

James, W. (1884). What is an emotion? *Mind, 9*, 188-205.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697-720.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: the endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5, 193-206.

Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263-291.

Kahneman, D. & Tversky, A. (1982). The psychology of preferences. *Scientific American*, 246(1), 160-173.

Kahneman, D. & Tversky, A. (2000). *Choices, Values, and Frames*. New York: Cambridge University Press.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics, 112,* 375-405.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience, 25*(19), 4806-4812.

Koehler, D. J., Brenner, L. A., & Tversky, A. (1997). The enhancement effect in probability judgment. *Journal of Behavioral Decision Making*, 10, 293-313.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature,* 435, 673-676.

Kreps, D. M. (1990). *A course in microeconomic theory*. Princeton: Princeton University Press.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion, 14*, 473-493.

Li, Y. & Bayliss, D. A. (1998). Presynaptic inhibition by 5-HT_{1B} receptors of glutamatergic synaptic inputs onto serotonergic caudal raphe neurones in rat. *Journal of Physiology*, 510(1), 121-134.

Litt, A., Eliasmith, C., & Thagard, P. (2006). Why losses loom larger than gains: modeling neural mechanisms of cognitive-affective interaction. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 495-500). Mahwah, NJ: Lawrence Erlbaum.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*, 267-286.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 585-589.

McClure, S. M., York, M. K. & Montague, P. R. (2004). The neural substrates of reward processing in humans: the modern role of fMRI. *The Neuroscientist*, 10(3), 260-268.

McGraw, A. P., Mellers, B. A., & Tetlock, P. E. (2005). Expectations and emotions of Olympic athletes. *Journal of Experimental Social Psychology*, 41, 438-446.

Mellers, B. A. (2000). Choice and the relative pleasure of consequences. *Psychological Bulletin*, 126, 910-924.

Mellers, B. A. & McGraw, A. P. (2001). Anticipated emotions as guides to choice. *Current Directions in Psychological Science*, 10, 210-214.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: emotional reactions to the outcomes of risky options. *Psychological Science*, 8, 423-429.

Mellers, B. A., Schwartz, A., & Ritov, I. (1997). Emotion-based choice. *Journal of Experimental Psychology: General*, 128, 332-345.

Mobini, S., Chiang, T. J., Ho, M. Y., Bradshaw, C. M. & Szabadi, E. (2000). Effects of central 5-hydroxytryptamine depletion on sensitivity to delayed and probabilistic reinforcement. *Psychopharmacology (Berl.),* 152, 390-397.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*(10), 799-809.

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

Oatley, K. (1992). *Best laid schemes: The psychology of emotions*. Cambridge: Cambridge University Press.

Owen, A. M. (1997). Cognitive planning in humans: neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*, 53, 431-450.

Padoa-Schioppa, C. & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature,* 441, 223-226.

Paré, D. & Gaudreau, H. (1996). Projection cells and interneurons of the lateral and basolateral amygdala: distinct firing patterns and differential relation to theta and delta rhythms in conscious cats. *Journal of Neuroscience,* 16(10), 3334-3350.

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of Neurophysiology*, 90, 1392-1407.

Roitman, J. D., Brannon, E. M., & Platt, M. L. (2007). Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS Biol*ogy, 5(8), e208 doi:10.1371/journal.pbio.0050208

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, 10, 284-294.

Rottenstreich, Y. & Hsee, C. K. (2001). Money, kisses and electric shocks: on the affective psychology of risk. *Psychological Science*, 12, 185-190.

Rottenstreich, Y., & Shu, S. (2004). The connections between affect and decision making: Nine resulting phenomena. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 444-463). Oxford: Blackwell.

Sanfey, A. G., Loewenstein, G., McClure, S. M., & Cohen, J. D. (2006). Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Sciences, 10*(3), 108-116.

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, 1, 199-207.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80, 1-27.

Shiv, B., Bechara, A., Levin, I., Alba, J. W., Bettman, J. R., Dubé, L., Isen, A., Mellers, B., Smidts, A., Grant, S. J. & McGraw, A. P. (2005). Decision neuroscience. *Marketing Letters*, 16, 375-386.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 397-420). Cambridge: Cambridge University Press.

Soubrié, P. (1986). Reconciling the role of central serotonin neurons in human and animal behavior. *Behavioral and Brain Sciences*, 9, 319-335.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks*, 15, 523-533.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

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

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811), 515-518.

Trepel, C., Fox, C. R., & Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive Brain Research, 23*(1), 34-50.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, 24(5), 295-300.

Tversky, A. & Kahneman, D. (1991). Loss aversion in riskless choice: a reference dependent model. *Quarterly Journal of Economics*, 106, 1039-1061.

Tversky, A. & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, 251-278.

Tversky, A. & Kahneman, D. (1984). Choices, values, and frames. *American Psychologist*, 39, 341-350.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.

Vertes, R. P. (1991). A PHA-L analysis of ascending projections of the dorsal raphe nucleus in the rat. *Journal of Comparative Neurology*, 313(4), 643-668.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Wagar, B. M., & Thagard, P. (2004). Spiking Phineas Gage: A neurocomputational theory of cognitive-affective integration in decision making. *Psychological Review, 111,* 67-79.

Wallis, J. D. & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, 18(7), 2069-2081.

Weber, B., Aholt, A., Neuhaus, C., Trautner, P., Elger, C. E., & Teichert, T. (2007). Neural evidence for reference-dependence in real-market-transactions. *NeuroImage,* 35, 441-447.