Single neuron correlates of learning, value, and decision in the human brain

Thesis by
Tomas Gallo Aquino

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended January 20th 2022

Tomas Gallo Aquino
ORCID: 0000-0002-6944-1053

ACKNOWLEDGEMENTS

The Brazilian educator Paulo Freire once stated: “Liberating education consists in acts of cognition, not transferals of information”. Indeed, learning can be one of the most transformative acts in a person’s life, and it is deeply dependent on one’s past experiences, on how one interacts with one’s environment, and, ultimately, in the cognitive implementation of learning strategies by the brain. Providing even a tiny contribution in the collective endeavor to understand these processes fills me with an enormous sense of purpose and realization, and I want to thank all the people who helped me all the way here.

Firstly, I want to thank my advisor, John O’Doherty, and my co-advisor, Ueli Rutishauser. I believe that their joint contribution to my education was complementary in the best way possible. I thank John for creating a friendly and stimulating work environment where all of us can achieve their potential, for patiently imparting to me his deep knowledge of the field, and for helping me hone my sometimes far-fetched ideas. I thank Ueli for all of our stimulating discussions on science and career matters, and for all the hands-on mentoring which was crucial in my development. Both of them have been true role models in the extremely challenging job of running a lab, including the diligence to keep the lights turned on, the ethics to manage a diverse group of students and postdocs, and the ever present urge to learn, grow and take risks in the pursuit of scientific goals.

Before coming to Caltech to study value learning and decision making I was working on theoretical models of time perception, as a visiting student at Universidade Federal do ABC (UFABC), while undertaking the Molecular Sciences "major" at the University of São Paulo (USP). I believe that formative period was crucial for my development, and I want to highlight some people who taught me valuable lessons at that time. My former advisors at UFABC, Marcelo Reyes and Raphael de Camargo, provided me with my first foray into the computational modeling of neural phenomena, and are a major reason I have joined the field of neuroscience at all. And one of my teachers while at USP, Renato Vicente, who perhaps has been more influential than he even realizes in the construction of my scientific interests, at a crucial time in my intellectual development.

I want to thank all the members of the O’Doherty and Rutishauser labs, who have contributed to my personal and professional growth in many ways. Being surrounded

by extremely bright and motivated people on a daily basis has been an immense privilege which I never take for granted, and they have made this experience all the more rewarding. In special I want to single out some people: Jeff Cockburn, who has been a close mentor from the very beginning and taught me the value of a good model and of the simple things in life; Juri Minxha, who had a decisive participation in the beginning of my graduate school years, including mentoring me on electrophysiology and working with patients in general; Logan Cross, who is also in my Computation and Neural Systems cohort at Caltech, and has been an inspiring colleague all along the way.

I also want to thank the members of my committee, John Allman and Ralph Adolphs, whose feedback and support have elevated the quality of my work. Also, I want to thank all the friends I made through all these years at Caltech and USP. I want to single out Lucas Meirelles, John Ciemniecki, Giovani Tomaleri, Rafael Ruggiero, Caique Ronqui, Brunno Reimann, and Igor Zenker: you all inspire me and taught me a lot. Also from the Caltech community, I must thank all the International Student Program officers, who make it so much easier to transition to life in the United States and at Caltech — I owe a lot to them. Additionally, I must thank the epilepsy patients who generously agreed to participate in our work during a challenging time in their lives, as well as the team of nurses and physicians at the Cedars-Sinai Medical Center, whose sense of duty inspires my awe and respect. Without them, none of this would be possible.

Lastly, I want to thank some very important people in my life: my parents, Sérgio and Eliane, who gave me everything and made me who I am, without ever asking for anything in return; The Lin family, who offered me their generosity and hospitality in these troubled times; and Kelly, who brings out the best in me, motivating me to keep following my dreams, while supporting me through my worst times. Thank you for sharing all of these moments with me and inspiring me to a more whimsical life every day.

Despite all the positive and life changing aspects of my doctorate experience, it is impossible to conclude this journey without a bittersweet feeling. The distance from close friends and family, which would already be difficult to navigate on its own, was amplified by the COVID-19 pandemic, which deeply affected international scholars. Beyond this short-term existential threat, it has been desolating to powerlessly watch from afar as ill-intentioned actors systematically erode the scientific and cultural institutions of Brazil and deliberately promote the distrust of science as political

strategy. Yet I am one of the lucky few who were given the conditions to jump through countless hurdles just to make it here, including pointless standardized tests, prohibitive application fees, and the invisible burden of lacking an international contact network in top institutions, to name a few. The reality is that countless talented and motivated individuals in Brazil are having their dreams of pursuing a scientific career denied or are being forced to leave in search of better opportunities. Both in Brazil and in the US, whether we are willing to acknowledge it or not, there is an active dispute for the hearts and minds of the public in which the Enlightenment ideals and worldview are in contention. Therefore, I believe we must be unapologetically proselytizing in our scientific practice and communication.

ABSTRACT

In this thesis, I present several new results on how the human brain performs value-based learning and decision-making, leveraging rare single neuron recordings from epilepsy patients in vmPFC, preSMA, dACC, amygdala, and hippocampus, as well as reinforcement learning models of behavior. With a probabilistic gambling task we determined that human preSMA neurons integrate computational components of stimulus value such as expected values, uncertainty, and novelty, to encode an utility value and, subsequently, decisions themselves. Additionally, we found that post-decision related encoding of variables for the chosen option was more widely distributed and especially prominent in vmPFC. Additionally, with a Pavlovian conditioning task we found evidence of stimulus-stimulus associations in vmPFC, while both vmPFC and amygdala performed predictive value coding, establishing direct evidence for model-based Pavlovian conditioning in human vmPFC neurons. Finally, in a Pavlovian observational learning paradigm, we found a significant proportion of amygdala neurons whose activity correlated with both expected rewards for oneself and others, and in tracking outcome values received by oneself or other agents, further establishing amygdala as an important center in social cognition. Taken together, our findings expand our understanding of the role of several human cortical brain regions in creating and updating value representations which are leveraged during decision-making.

PUBLISHED CONTENT AND CONTRIBUTIONS

Aquino, Tomas G., Jeffrey Cockburn, Adam N. Mamelak, Ueli Rutishauser, and John P. O'Doherty (2021). "Neurons in human pre-supplementary motor area encode key computations for value-based choice". In: *bioRxiv*. DOI: 10.1101/2021.10.27.466000.

T.G.A., J.C., U.R., and J.P.O. designed the study. T.G.A. performed the experiments. T.G.A. and J.C. analyzed the data. T.G.A., J.C., A.N.M. U.R., and J.P.O. wrote the paper. A.N.M. performed surgery and supervised clinical work.

Aquino, Tomas G, Juri Minxha, Simon Dunne, Ian B Ross, Adam N Mamelak, Ueli Rutishauser, and John P O'Doherty (2020). "Value-related neuronal responses in the human amygdala during observational learning". In: *Journal of Neuroscience*. DOI: 10.1523/JNEUROSCI.2897-19.2020.

T.G.A. performed spike data pre-processing and all subsequent data analyses, and wrote the manuscript; J.M. collected data, performed spike data pre-processing and wrote the manuscript; S.D. created the original task design; I.B.R. performed patient surgery; A.N.M. performed patient surgery; U.R. and J.O.D. both conceived and directed the project, and wrote the manuscript.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

AMY. Amygdala.

BLA. Basolateral amygdala.

BOLD. Blood-oxygen-level-dependent.

dACC. Dorsal anterior cingulate cortex.

dlPFC. Dorsolateral prefrontal cortex.

fMRI. Functional magnetic resonance imaging.

HIP. Hippocampus.

iEEG. Intracranial electroencephalography.

LIP. Lateral intraparietal area.

OFC. Orbitofrontal cortex.

PPC. Posterior parietal cortex.

preSMA. Pre-supplementary motor area.

SEM. Standard error of the mean.

vmPFC. Ventromedial prefrontal cortex.

VTA. Ventral tegmental area.

Chapter 1

GENERAL INTRODUCTION

1.1 Overview

The ability to learn from the environment and improve future decisions based on information acquired in the past is a fundamental adaptive characteristic of humans and other animals. From learning simple associations between stimuli and outcomes to creating complex cognitive maps that elicit optimal action planning in uncertain environments, our experiences are constantly shaped by how our brains represent, process, store, and utilize multidimensional information.

Consider, for instance, the problem of choosing where to order food on a given night. Perhaps you are deciding between a familiar restaurant, which you have visited several times, and a new restaurant in town. Think about all the processes that need to occur in your brain between seeing the possible options on your favorite food ordering app and finally pressing the button to submit your order. Among the many variables which determine your final decision, including your preference for a certain cuisine, or your current cravings, it is reasonable to assume that you would compute an estimate of the overall value of each possible choice in order to compare them. Importantly, your past experiences should play a role in computing an estimate for the quality of the food from each restaurant, and such estimates are likely more accurate for the familiar restaurant than for the novel one, on the basis of having sampled the former. Still, even though you might have enjoyed the familiar restaurant in the past, you might be compelled to explore the novel option out of curiosity, which ultimately will help you acquire information about it and reduce the degree of uncertainty in estimating its quality. This is, in essence, the explore-exploit dilemma (Sutton and Barto, 2018), which is one example of how value estimates based on one's prior experiences might be integrated with other stimulus features to inform decision-making. In the scope of this thesis, it is of particular interest how single neurons in the human brain participate in different stages of this process, from representing values and other stimulus features to producing a final decision output.

At a more fundamental level, another aspect of value learning which has not yet been fully mapped in human neurons is Pavlovian conditioning, which is defined

by learning to predict potentially rewarding or punishing outcomes from initially neutral stimuli. On the one hand, nearly a century of behavioral and neural studies have significantly advanced our understanding of which brain areas are involved in learning and utilizing associations between stimuli and outcomes, in humans and other animals (Pavlov, 1927; R. A. Rescorla, 1988; Sharpe and Schoenbaum, 2016; O'Doherty, Cockburn, and Pauli, 2017). On the other hand, the computational implementation of this process by neural populations is yet to be understood in its full complexity.

Take, for example, the issue of sensory preconditioning in Pavlovian learning (Brogden, 1947). When a group of dogs learned to associate a sound tone and a light pulse, and subsequently learned that either the tone or the light were predictive of a shock, the other stimulus would immediately start eliciting a fear response, even though it had never been presented with the shock itself. While it has been established that this form of higher-order learning also occurs in humans (White and Davey, 1989), the most ubiquitous framework used to explain classical conditioning behavior have been model-free (MF) models (R. Rescorla and Wagner, 1972; Sutton, 1988), which rely exclusively on experiencing outcomes to update their cached value predictions for stimuli, and are thus insufficient to capture phenomena such as sensory preconditioning. For this reason we investigated the creation of cognitive maps (Tolman, 1948) during Pavlovian conditioning, which allow for a more flexible learning approach than the MF framework. Crucially, these models encode a transition structure between task states, beyond caching stimulus values (O'Doherty, Cockburn, and Pauli, 2017). We also investigated their implementation by human neural populations in the context of higher-order Pavlovian conditioning, leveraging recordings in areas which have been established to participate in this form of learning (Prévost, McNamee, et al., 2013; Pauli, Gentile, et al., 2019), including ventromedial prefrontal cortex and amygdala.

As we expand our understanding of how neural populations implement learning strategies during instrumental and Pavlovian conditioning, new sets of questions arise. One intriguing research direction is the interface between value learning and social cognition, as valuable insight can be obtained from studying how agents learn from observing others (Carcea and Froemke, 2019). From songbirds learning to vocalize by hearing their peers (Mooney, 2014) to rodents learning about threats by observing fear responses in their partners, utilizing a circuitry involving ACC and amygdala (Allsop et al., 2018), there are clear adaptive advantages to learning from

one's conspecifics rather than in isolation.

One theory postulates that observational learning stems partially from a *mirror neuron* system, which was initially discovered in the monkey premotor cortex (Rizzolatti et al., 1996) and postulated to be observed in the human brain through neuroimaging (Iacoboni et al., 1999). These neurons were found to activate when the monkey performed certain hand movements and, crucially, when it saw a human perform the same motions. Above and beyond encoding movements performed by others, humans are also known to recognize emotions in others — in facial expressions for instance (Olsson and Phelps, 2007). This information can then be utilized to drive one's own emotional states accordingly, relying on a corticoamygdalar circuitry.

Given the amygdala's role in social cognition, as well as in value learning, in tandem with OFC/vmPFC (Sharpe and Schoenbaum, 2016; O'Doherty, Cockburn, and Pauli, 2017), we leveraged single neuron recordings in human amygdala to investigate how expected values and outcomes were encoded by neural populations during an observational reinforcement learning task, contrasting such neural activity patterns with self-experienced learning.

Overall, the present dissertation provides a number of new insights from the fruitful interaction between reinforcement learning theory and electrophysiological recordings in humans, made possible by invasive monitoring in refractory epilepsy patients. In special, I will focus on the role of human preSMA and vmPFC in representing and integrating several features of value and producing decision outputs when acting under uncertainty. Additionally, I will present new data on the role of vmPFC and amygdala in creating cognitive maps for Pavlovian conditioning, and explore data we collected on the amygdala in the context of social cognition.

1.2 Neural correlates of value learning and decision making

Representing values and expectations in neural activity

To fully motivate our research on representations of value-based learning in human neurons, I will provide an overview of electrophysiology and value studies, including studies in humans and other animals. I will also pay special attention to the anatomy of the human prefrontal cortex, which will provide a helpful foundation for all of our studies, presented subsequently.

One of the first questions which must be explored is whether the activity of individual neurons can be used by organisms to represent values at all. In this context, we will broadly define *value* as a quantifiable quality assigned to stimuli or actions by

agents, which reflects their subjective preferences and is used to drive decisions toward options or actions of higher value.

In a landmark study, the activity of some neurons in the monkey orbitofrontal cortex (OFC) was found to correlate with the economic value of offered and chosen goods during a choice task (Padoa-Schioppa and Assad, 2006). Crucially, this correlation held true regardless of spatial factors or motor responses. Similarly, recordings in monkeys performing a reward preference task revealed that while dorsolateral prefrontal cortex (dlPFC) neurons encoded rewards and monkeys' forthcoming responses, OFC neurons tended to only encode reward values (J. D. Wallis and E. K. Miller, 2003). Taken together, these results indicate a role for OFC in encoding abstract value-related responses regardless of subsequent performed behaviors.

Several electrophysiological studies revealed that, in fact, value coding in OFC occurs as part of a broader neural circuitry involving the basolateral and centromedial nuclei of the amygdala (Schoenbaum, Chiba, and Gallagher, 1998; Schoenbaum, Setlow, et al., 2003; Salzman and Fusi, 2010; Sharpe and Schoenbaum, 2016), which is known to encode economic values (Jenison et al., 2011) and behavioral choices (Grabenhorst, Hernádi, and Schultz, 2012). These areas are deeply connected anatomically, bidirectionally, (Aggleton, Burton, and Passingham, 1980; Ghashghaei, Hilgetag, and Barbas, 2007), and activity in amygdalar neurons had previously been established as a substrate for acquisition of Pavlovian responses (Applegate et al., 1982), specifically encoding expected values acquired during pleasant or aversive conditioning (Belova, Paton, Morrison, et al., 2007). Additionally, recent studies have provided a more detailed understanding of the relationship between amygdala and vmPFC/OFC. Initial evidence suggested that specific lesions to OFC but not BLA made rats unable to decrease conditioned responses after stimulus devaluation, suggesting that while BLA played a role in encoding expected values of stimuli (Schoenbaum, Chiba, and Gallagher, 1998), it did not play a role in updating these representations with new information or in adapting behavior accordingly, a role which might pertain to OFC (Pickens et al., 2003). More recently, through chemogenetic and optogenetic manipulations it was determined that projections from lateral OFC to BLA were necessary and sufficient to encode reward values in memory, while projections from medial OFC to BLA were necessary and sufficient to retrieve these values from memory (Malvaez et al., 2019).

In the human brain, OFC was found to signal outcome identity, during the presentation of a stimulus which was learned to predict that identity (Howard et al.,

2015). More generally, neuroimaging studies have found responses in OFC, ventral striatum, and amygdala correlating with outcome predictive stimuli, both for appetitive and aversive outcomes (Gottfried, O'Doherty, and Dolan, 2002; Gottfried, O'Doherty, and Dolan, 2003; Tobler et al., 2006), while activity in the OFC was also found to correlate to outcome values themselves (De Araujo, Rolls, et al., 2003; De Araujo, Kringelbach, et al., 2003; Kahnt et al., 2010).

The ventromedial prefrontal cortex (vmPFC) is another prefrontal area which plays a similarly central role in value coding in the brain (O'Doherty, Cockburn, and Pauli, 2017). Anatomically, it is adjacent to OFC and is bilaterally connected to it, as well as amygdala (Joseph L Price, 1999). In fMRI, blood-oxygen-level-dependent (BOLD) signal in vmPFC is predictive of the value of goods in a willingness-to-pay paradigm, distinguishing appetitive goods from aversive ones (Plassmann, O'Doherty, and Rangel, 2010), and also correlates with values across different categories of goods, including money, food and non-food consumer items (Chib et al., 2009; Levy and Glimcher, 2012). These findings suggested the existence of a common currency code in vmPFC, which would in theory allow for different items to be compared in value before a decision. Additionally, more ventral areas of vmPFC were later found to encode category-dependent value codes (across food or consumer items), while a more dorsal region was found to contain a category-independent value code, (McNamee, Rangel, and O'Doherty, 2013). Further work established that the value of different nutritional attributes of food (fat, protein, carbohydrates and vitamins) was represented laterally in OFC, while an integrated overall value was computed in a medial OFC region compatible with vmPFC (Suzuki, Cross, and O'Doherty, 2017). Taken together, these findings offer support to the hypothesis of a common currency computation in vmPFC stemming from independent components of value partially computed elsewhere.

Updating value representations

Having established that a corticoamygdalar circuitry is involved in encoding abstract values, including the expectation of rewards and punishments, it is fitting to ask how these learned representations are acquired in the first place. In other words, how is new information utilized to update an organism's model of the world to generate new expectations? A compelling candidate for a reward learning signal is the difference between one's expectations and the actual outcomes received, or the amount of surprise experienced by an organism (R. Rescorla and Wagner, 1972). In reinforcement learning (RL) theory, this value is known as reward prediction error (RPE),

and a seminal study in rodents has proposed that this quantity is represented in the phasic activity of midbrain dopaminergic neurons (Schultz, Dayan, and Montague, 1997), which project to striatum but also several other cortical targets. Congruently, neuroimaging results have reported correlations with RPE in the human striatum and dopaminergic midbrain (O'Doherty, Dayan, et al., 2003; D'Ardenne et al., 2008). This hypothesis has been since expanded to account for representing entire distributions of possible rewards, above and beyond expected values alone (Bellemare, Dabney, and Munos, 2017; Dabney, Rowland, et al., 2018). This framework, known as distributional RL, provides an accurate account of how neural populations represent RPEs to generate expectation distributions (Dabney, Kurth-Nelson, et al., 2020), potentially providing a fruitful new direction for this research field.

Neural encoding of value-based decisions

Once an organism has determined the values of possible actions, a series of processes must take place to convert these values into actual decisions, encompassing value comparisons but also a consideration of an organism's own confidence about its value estimates (i.e. metacognition). Important insight into such processes has been obtained from studying perceptual decision making tasks, in which the values of motor decisions and the level of confidence about them can be systematically manipulated. For instance, neurons in the lateral intraparietal area (LIP) of monkeys, an area of posterior parietal cortex (PPC) which is known to encode directional saccade movement planning (Gnadt and Andersen, 1988), were also found to encode the expected reward of performing an eye movement action (Platt and Glimcher, 1999), tentatively suggesting one neural substrate for integrating value information with the execution of motor commands. Indeed, later work has suggested a mechanism for how this integration occurs, using a discrimination of motion direction task, with a reward manipulation (Rorie et al., 2010). In this task, monkeys had to perform saccades to targets positioned on either side of the screen, to report which one of two possible sides was the perceived direction of motion for most of the randomly moving dots on the screen. These dots had varying degrees of motion coherence, which made it easier or harder to detect the correct direction. Crucially, each of the possible targets was assigned a reward magnitude (low or high) in case it was correctly targeted. With this manipulation, is was determined that LIP neurons were influenced concurrently by the dots' motion coherence but also both reward values, a pattern which was computationally explained by using a reward biased drift-diffusion model. Indeed, drift-diffusion models have been used successfully

to explain the relationship between decisions, reaction times, and value-induced biases in a variety of contexts (Gold and Shadlen, 2007; Milosavljevic et al., 2010; Krajbich and Rangel, 2011).

Another area of particular interest in this context is dlPFC, since it anatomically bridges the value-coding prefrontal areas and the premotor areas responsible for motor planning, including preSMA (Luppino et al., 1993), as I will discuss further. In monkeys, neurons from this brain area integrated visual sensory signals with motor planning, reflecting the monkey's subsequent gaze shifts (Kim and Shadlen, 1999), represented the past history of decisions and outcomes in a multi-agent decision-making task (Barraclough, Conroy, and D. Lee, 2004), and encoded rewards and monkeys' forthcoming responses in a reward preference task (J. D. Wallis and E. K. Miller, 2003). In human neuroimaging, dlPFC has been suggested as a potential locus for the integration of sensory information for the purpose of decision making (Heekeren et al., 2004). Additionally, dlPFC activity has been found to correlate with the variability of value attributes when creating an integrated value for multi-attribute items (Kahnt et al., 2011), which could be interpreted as a metacognitive measure of ambiguity or difficulty in integration. Furthermore, dlPFC has also been implicated in cognitive control during value-based tasks, as activity in this brain area correlated with choosing delayed rewards in a delayed gratification task (Hare, Hakimi, and Rangel, 2014).

Mapping reinforcement learning in the brain

So far I have tangentially discussed the matter of using reinforcement learning models to capture certain computational aspects of value learning and decision making which can, in turn, be mapped to brain activity in humans and other animals. Given our interest in exploring learning phenomena in human electrophysiology, and the widespread adoption of this class of models (O'Doherty, Cockburn, and Pauli, 2017), I will now discuss them in more depth.

As I briefly introduced, one of the most influential ideas in neural reinforcement learning is the usage of a learning signal based on the discrepancy between expectations and outcomes, or reward prediction errors. One of the simplest yet most fundamental applications of this idea is in temporal difference (TD) learning models, which iteratively incorporate new evidence through RPEs weighted by a learning rate parameter. To support our discussion of this and other models, I will first remind the reader of a few fundamental definitions pertinent to reinforcement

learning (Sutton and Barto, 2018).

In a general formulation of this problem, an agent interacts with the environment over a series of time steps $t = 0, 1, 2, ..., T$ by performing actions A_t from a possible action set. As a consequence of its action, the state of the environment changes from S_t to S_{t+1} according to some probabilistic structure, over a set of possible states. Additionally, depending on the state S_t visited by the agent, it might receive a reward value R_t, according to another probabilistic structure. Generally speaking, the objective of an agent, and by extension, of a training algorithm, is to maximize the expected sum of rewards over the course of time, which we define as the return $G_t := R_{t+1} + R_{t+2} + ... + R_T$. To account for the fact that the total time T might be infinite, over a continuing decision problem, an additional factor which is typically added to this definition is *discounting*, which reflects the idea that rewards obtained in a distant future are valued less than rewards obtained immediately: $G_t := R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$, with a discount factor $0 \leq \gamma \leq 1$.

Given the actions available to the agent and possible environment states, we can define a *policy* π, which is a mapping from states s to the probability of selecting each possible action a: $\pi(a|s)$. One interesting problem is therefore how to select the best possible policy in order to maximize rewards, given the probabilistic structure of the environment. Having defined a policy π, we can propose a *value function* $v_\pi(s)$ that we are trying to optimize, by taking the expected return starting in state s, assuming the agent will follow the policy π: $v_\pi(s) = E_\pi[G_t|S_t = s]$.

In many real applications, the agent does not have perfect knowledge of what the value function is, and, therefore, it might be necessary to operate with an estimate $V(S_t)$ for $v_\pi(s)$, which is iteratively updated as the agent interacts with the environment and gathers information. Indeed, TD-learning is one example of such an approximative approach, which updates the value estimate $V(S_t)$ at every time step, using the reward and value estimates observed at time step $t + 1$:

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{1.1}$$

Unpacking this equation, at every time step $V(S_t)$ is changed from its original value by adding an error factor corresponding to the difference between the current value estimate and an improved value estimate, comprised by the observed reward at time $t + 1$ plus a discounted estimated of the value at $t + 1$. This factor is weighted by a learning rate parameter α before being utilized to update value estimates. This

weighted quantity is referred to TD-error and corresponds to the reward prediction error value which has been found to map to neural activity in the dopaminergic midbrain as previously mentioned.

TD-learning algorithms are one simple but effective method to optimize an agent's value estimates. Its computational simplicity is a double-edged sword, in that it requires few resources to implement, either in a biological or artificial agent, but at the cost of flexibility which could lead to better decisions. Specifically, TD-learning ignores the transition structure which dictates the probability $p(S_{t+1} = s'|S_t = s)$ of transitioning from some state s to another state s', meaning that states cannot have their values updated unless they are directly experienced, forgoing the ability of inferring updated state values from the full transition structure of the environment. Models which exhibit this characteristic are referred to in the reinforcement learning literature as model-free.

The alternative to model-free algorithms is the model-based class of learning algorithms, which incorporates the idea of planning possible future paths by encoding and utilizing a transition probability model for the environment. The extra flexibility afforded by these methods comes at the price of additional computational complexity, however. One simple example of a model-based learning algorithm is the FORWARD learner (Gläscher et al., 2010), described as follows. The crucial component of this algorithm is a transition matrix $T(s, a, s')$ which maps the probability of transitioning from state s to state s' if action a is performed. Assuming this matrix is known, one may recursively compute the value of a state-action pair $Q(s, a)$, evaluating the following equation:

$$Q(s, a) = \sum_{s'} T(s, a, s') \cdot (R(s') + \arg\max_{a'} Q(s', a')) \tag{1.2}$$

The recursive computation into future states an actions is done considering the agent greedily selects the action which maximizes the state-action value. This computation is tractable for actual applications for a small number of possible states, assuming that the structure of the problem at hand leads to some final state after which the learning event is over. In the field of dynamic programming, this class of recursive value evaluation equations is collectively known as the *Bellman equation*.

To actually learn the transition matrix $T(s, a, s')$, we can use a learning signal which is the discrepancy between the probability of arriving into a state and the agent's

encoded expectation, given the performed action. This learning signal is known as the state prediction error (SPE), defined as follows:

$$SPE = 1 - T(s, a, s') \tag{1.3}$$

The transition probability matrix can then be updated accordingly with a learning rate parameter η:

$$T(s, a, s') \leftarrow T(s, a, s') + \eta SPE \tag{1.4}$$

To make sure that the transition probabilities remain normalized and still represent actual probabilities, transition probabilities to states s'' other than s' are also updated: $T(s, a, s'') \leftarrow T(s, a, s'') \cdot (1 - \eta)$.

A learner adopting the FORWARD algorithm is able, for example, to immediately adapt its course of action if states further into the future are suddenly devalued, since it is able to estimate which actions might eventually lead to that newly undesirable state.

A long standing hypothesis in psychology and neuroeconomics is that humans and other animals make decisions according to the output of at least two systems, one of which is slow, but flexible and goal-directed, while the other one is fast, but inflexible and habitual (Balleine and Dickinson, 1998; Kahneman, 2011). In theory, it would be advantageous to have both systems in place, depending on the organism's current environmental demands. While different formulations of this idea have been proposed, given the trade-off between the complexity and the flexibility afforded by model-free and model-based learning algorithms, these two algorithm classes have been proposed as candidates to model habitual and goal-directed learning, respectively (Daw, Niv, and Dayan, 2005).

One question which naturally arises from the existence of multiple behavioral controllers is how the brain arbitrates between different controllers, and which factors lead each controller to increase or decrease its influence over behavior. One proposed solution is for the accuracy of each controller's predictions to serve as an arbitration signal, in which case more precise controllers would exert a larger influence over final behavioral outputs (Daw, Niv, and Dayan, 2005; O'Doherty et al., 2021). A proxy for each controller's accuracy which is readily available is its own prediction error signal (RPE for model-free learners and SPE for model-based learners), which

indicates the extent to which experiences violate the algorithm's expectations. Indeed, the existence of a neural arbitrator between these two systems was supported by neuroimaging evidence in which inferior lateral prefrontal cortex and frontopolar cortex encoded model-free and model-based reliability signals, as well as their direct comparison, indicating that these regions encode sufficient information for controller arbitration to take place. A proposed mechanism for the neural arbitrator is that it directly down-regulates the model-free learning system, whenever the model-based reliability was relatively higher than model-free reliability (S. W. Lee, Shimojo, and O'Doherty, 2014).

The idea that distinct controllers can produce competing outputs which contribute to behavior is not exclusive to the model-free/model-based distinction, though this is one of the examples that have been studied most extensively (O'Doherty et al., 2021). For example, in the context of social cognition, two possible mechanisms that can be arbitrated for observational learning are imitation or emulation (Charpentier, Iigaya, and O'Doherty, 2020). In imitation learning, an agent acts by copying an observed agent's actions, with the assumption that performing the same actions may lead to better outcomes. In emulation learning, however, an organism attempts to infer the observed agent's goals to inform its own decision making strategy, instead of simply copying the observed agent's actions. In similar fashion to the model-based/model-free dichotomy, a reliability-based arbitration explained subjects' behavior in an imitation vs. emulation task, and the reliability signal was found to correlate with activity in ventrolateral prefrontal cortex, ACC, and temporoparietal junction (TPJ).

Prefrontal cortex anatomy

To provide a common foundation for the discussion of all the work presented in this thesis, I will briefly review the anatomy of the primate prefrontal cortex, with a special focus on the connectivity between vmPFC, dACC, and preSMA, the main recurring prefrontal areas in our studies. For the purposes of this anatomical discussion I will use vmPFC as a shorthand for area 14 specifically, setting a distinction from areas 11, 12, and 13, which I will refer to as OFC. Additionally, I will focus mostly on area 24 when discussing dACC, as well as areas 9 and 46 for dlPFC. A summary of the discussed prefrontal anatomical connections is displayed on Fig. 1.1.

In non-human primates, anatomical tracing revealed that the vmPFC displays strong bidirectional connectivity with OFC, especially through area 13, as part of an

orbital-prefrontal network (Carmichael and J. Price, 1996; Joseph L Price, 1999). It also receives strong projections from amygdala and hippocampus (subiculum), two regions with known projections between each other (Rosene and Van Hoesen, 1977; Amaral, 1986). Furthermore, vmPFC displayed bidirectional connectivity with dlPFC, which has been partially supported by human fMRI tractography (Sallet et al., 2013).

The dlPFC was shown to receive projections from OFC, especially from area 13, while it sent heavy projections to OFC, from area 46 to areas 12 and 13, and from both areas 9 and 46 to vmPFC. (Carmichael and J. Price, 1996; Joseph L Price, 1999). It also received projections from the basal nucleus of the amygdala (Bozkurt et al., 2001). Through tracing, it was found that dlPFC also projects to dACC (Vogt and Pandya, 1987), congruently to human tractography results (Sallet et al., 2013). The dACC also receives important projections from OFC, amygdala, and hippocampus (subiculum and CA1) (Vogt and Pandya, 1987), while functional connectivity in humans has also been suggested with vmPFC (Du et al., 2020).

Also through anatomical tracing in non-human primates it was determined that preSMA has strong projections into SMA proper, and receives projections from dACC (Luppino et al., 1993). Additionally, the main prefrontal projections into preSMA originate in dlPFC, which situate the latter in a unique intermediary position between the value-coding prefrontal areas and the preSMA/SMA complex.

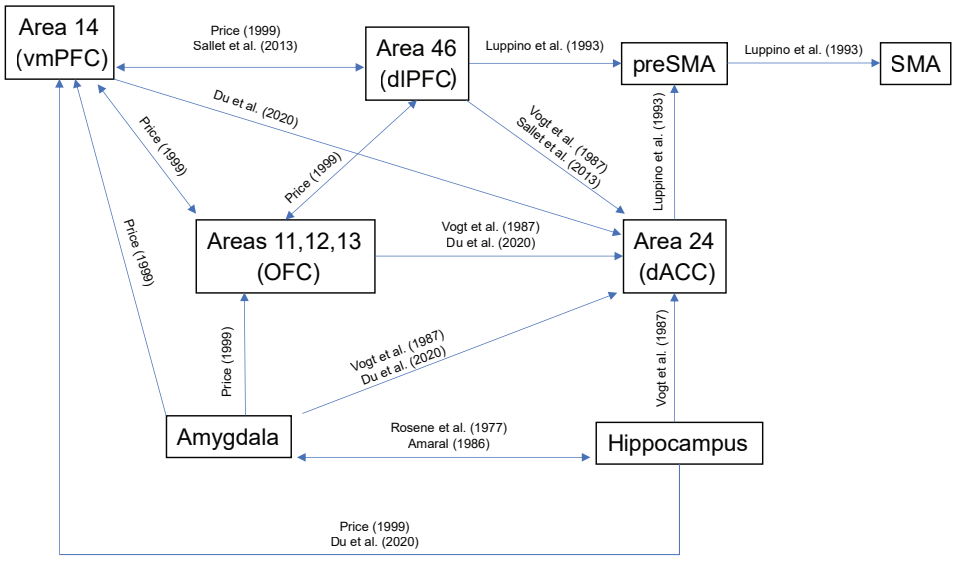

Figure 1.1: Simplified schematic of connectivity for primate prefrontal brain areas relevant for the present thesis, supported by anatomical tracing.

1.3 Overview of the thesis

Chapter 2 focuses on value-based feature integration performed by neurons in human preSMA, vmPFC and dACC, in the context of the explore-exploit dilemma. Chapter 3 focuses on correlates of model-based Pavlovian conditioning in human vmPFC and amygdala neurons. Chapter 4 focuses on value-based responses in human amygdala, contrasting learning through one's own experience with learning through observing others.

Chapter 2

NEURONS IN HUMAN PRE-SUPPLEMENTARY MOTOR AREA ENCODE KEY COMPUTATIONS FOR VALUE-BASED CHOICE

The following chapter is adapted from Aquino et al., 2021 and modified according to the Caltech Thesis format.

Aquino, TG, Cockburn, J, Mamelak, AN, Rutishauser, U, O'Doherty, JP, Neurons in human pre-supplementary motor area encode key computations for value-based choice, bioRxiv (2021), https://doi.org/10.1101/2021.10.27.466000

2.1 Abstract

Adaptive behavior in real-world environments demands that choices integrate over several variables, including the novelty of the options under consideration, their expected value, and uncertainty in value estimation. We recorded neurons from the human pre-supplementary motor area (preSMA), ventromedial prefrontal cortex (vmPFC) and dorsal anterior cingulate to probe how integration over decision variables occurs during decision-making. In contrast to the other areas, preSMA neurons not only represented separate pre-decision variables for each choice option, but also encoded an integrated utility signal and, subsequently, the decision itself. Conversely, post-decision related encoding of variables for the chosen option was more widely distributed and especially prominent in vmPFC. Our findings position the human preSMA as central to the implementation of value-based decisions.

2.2 Introduction

Humans and other animals can make decisions in a manner that maximizes the chance of obtaining rewards. Computational theories of decision-making suggest that doing so relies on a number of variables (Sutton and Barto, 2018). Most studied among these is the expected value (EV) associated with an option. By comparing options with varying EVs, it is possible to guide behavior toward higher expected future reward. However, in the real-world, the relationship between actions and their subsequent outcomes is often uncertain, and as such, one needs to consider not only the expected reward, but also its estimation uncertainty (Payzan-LeNestour and Bossaerts, 2012; Gershman, 2018). Another relevant feature is the novelty of an option — novel options can potentially provide new opportunities to gain

reward (Wittmann et al., 2008). These features can be utilized to resolve an often encountered dilemma in decision-making: whether to explore uncertain options that could yield richer reward, or to exploit an option with known rewards (Cohen, McClure, and A. J. Yu, 2007; R. C. Wilson et al., 2014).

How does the human brain represent the decision variables associated with the available options and how are they integrated to make a decision? One possibility is that neurons encode a utility signal that integrates over relevant decision variables for a given option and that this integrated utility is then used as an input to the decision process. Alternatively, these variables could be encoded in non-overlapping neuronal populations and be integrated at the population level to inform action selection.

Studies in rodents and non-human primates have reported neurons throughout the prefrontal cortex that correlate with EV (J. D. Wallis, 2007; Padoa-Schioppa and Cai, 2011; Grabenhorst and Rolls, 2011; Cai and Padoa-Schioppa, 2012; Strait, Blanchard, and Hayden, 2014; Rich and J. D. Wallis, 2016), uncertainty (Kepecs et al., 2008; O'Neill and Schultz, 2010; Grabenhorst, Báez-Mendoza, et al., 2019; Hirokawa et al., 2019) and novelty (Dias and Honey, 2002; M. Matsumoto, K. Matsumoto, and Tanaka, 2007; Bourgeois et al., 2012). Most human studies have been restricted to non-invasive methods such as functional magnetic resonance imaging (fMRI), revealing roles in value-based decision making for the vmPFC (Chib et al., 2009; Hare, Schultz, et al., 2011; Suzuki, Cross, and O'Doherty, 2017; Kobayashi and Hsu, 2019), dorsal anterior cingulate cortex (dACC) (Walton, Devlin, and M. F. Rushworth, 2004), and preSMA (Wunderlich, Rangel, and O'Doherty, 2009; Hare, Schultz, et al., 2011). Overall, these areas encode decision variables such as EV (Chib et al., 2009; Wunderlich, Rangel, and O'Doherty, 2009; Grabenhorst and Rolls, 2011; Hare, Schultz, et al., 2011; Kobayashi and Hsu, 2019), uncertainty (Badre et al., 2012; Kobayashi and Hsu, 2019; Trudel et al., 2021), and outcomes (Grabenhorst and Rolls, 2011; Vassena et al., 2014), while novelty related effects have also been found in the dopaminergic midbrain and striatum (Horvitz, Stewart, and Jacobs, 1997; Wittmann et al., 2008; Krebs et al., 2009; Kamiński et al., 2018). Some studies reported value computations in prefrontal cortex utilizing intracranial EEG (iEEG) from depth and grid electrodes (Saez et al., 2018; Domenech, Rheims, and Koechlin, 2020). While this approach affords greater temporal resolution than fMRI, iEEG reflects pooled activity across large numbers of neurons with a similar lack of spatial selectivity as fMRI. In particular, while previous studies (Wunderlich,

Rangel, and O'Doherty, 2009; Hare, Schultz, et al., 2011) demonstrated correlations with action value in supplementary motor cortex with fMRI, they do not show whether value-related signals precede decision-related signals and how these two signals interact.

We sought to determine how single neurons in these three brain areas are recruited during decision-related computations, to address whether these variables are integrated into a utility signal at the level of single neurons, and to probe how these signals might be utilized for informing choice. For this, we recorded single neurons in preSMA, dACC, and vmPFC while human patients with drug resistant epilepsy undergoing invasive electrophysiological monitoring performed a decision-making task specifically designed to dissociate EV, uncertainty and novelty. Additionally, we aimed to distinguish neurons that encode stimulus features and choice from those that evaluate the consequences of the decision. Finally, we could identify neurons encoding outcomes and prediction errors, to ascertain how these regions contribute to updating decision information following feedback at the neuronal level. Thus, this study afforded us an unparalleled opportunity to investigate the role of human prefrontal neurons across multiple stages of value-based decision-making: from the representation of individual decision variables, through to integration of these variables into a putative utility signal, up to choice and ultimately feedback.

2.3 Results

Task and behavior

We recorded 191 vmPFC, 137 preSMA and 108 dACC single neurons (436 total) in 22 sessions from 20 patients chronically implanted with hybrid macro/micro electrodes for epilepsy monitoring (Fig. 2.1A). Patients performed a two-armed bandit task (Cockburn et al., 2021) designed to separate the influence of EV, uncertainty and novelty on decision making, divided into 20 blocks consisting of 15 binary choices. On each trial, participants used a button box to choose between two uniquely identifiable bandits presented on the left or the right of the screen (Fig. 2.1B). Following a time delay, a feedback screen then announced the binary outcome (win/no win). The experimental design included two critical features. Firstly, participants were informed that the probability of each bandit delivering a reward was fixed for the duration of each block, but randomized across blocks. Secondly, both novel and familiar stimuli were systematically incorporated into the set from which options could be drawn during a block, resulting in pairs of bandits that varied in terms of EV, uncertainty, and novelty.

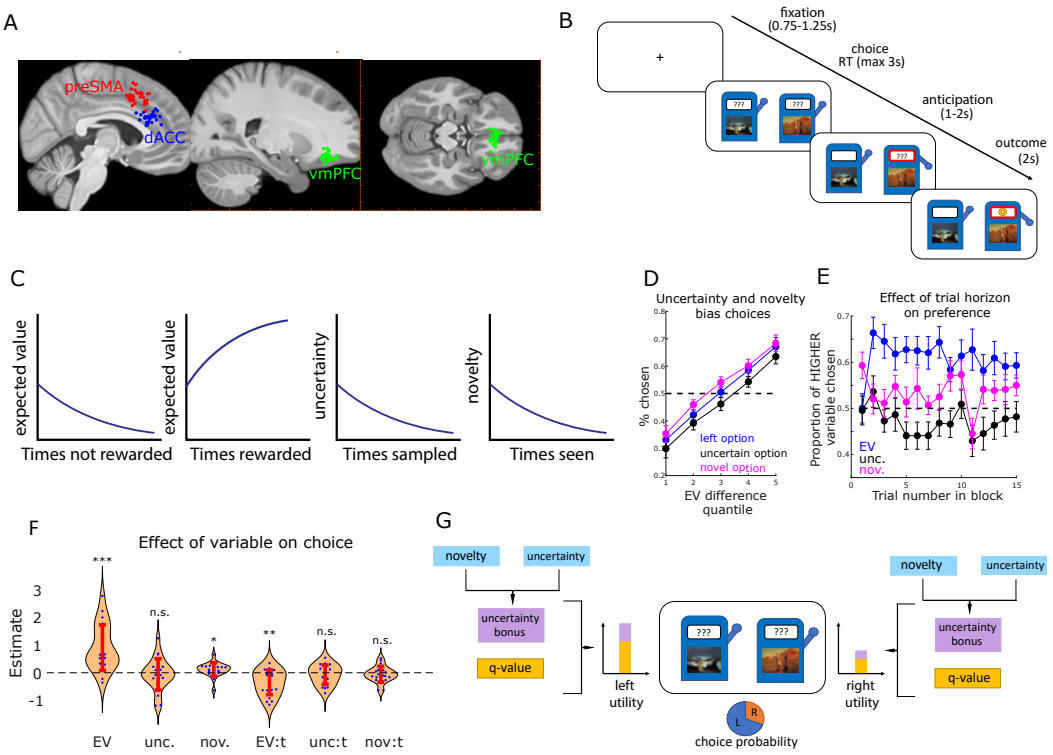

Figure 2.1: Exploration task, electrode positioning, and behavior. (a) Electrode positioning. Each dot indicates one electrode in one patient in preSMA (red), dACC (blue), or vmPFC (green). (b) Trials were structured in fixation, decision, anticipation, and feedback stages. (c) Schematic indicating how q-values, uncertainty, and novelty of stimuli vary as a function of the past history of rewards, sampling, and exposures. (d) Expected value correlates with choice, biased by novelty and uncertainty. Patients chose the left option (blue), the more uncertain option (black), or the more novel option (magenta) as a function of chosen minus unchosen expected value. (e) Proportion of trials in which patients chose the option with higher expected value (blue), uncertainty (black), or novelty (magenta), as a function of trial number. Dots and bars indicated means and standard errors, respectively. (f) Logistic regression coefficients for expected value, uncertainty, novelty, and interactions with trial number. Dots and bars indicate fits for each patient and standard error, respectively (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$, t-test). Positive values indicate seeking behavior. (g) fmUCB model. Novelty and uncertainty generate an uncertainty bonus, which composes utility along with q-values.

We first assess how EV, uncertainty and novelty related to behavior. Within each block of trials, EV and uncertainty were quantified as the average proportion of wins and total number of samples within a given block of trials respectively, while novelty was defined as the total number of times a stimulus had been shown across the entire experiment (see Fig. 2.1C). Uncertainty and novelty biased value-based decisions in

distinct directions: while on average patients preferred options with higher EVs over options with lower EVs ($p < 10^{-51}$, $T = 18.2$, linear regression), they sought them more often if they were also the more novel option, than if they were also the more uncertain option ($p < 0.01, T = 2.73$, two-sided t-test) (Fig. 2.1D). This was not the result of changing preferences over time because trial number did not correlate with how often patients sought the option with higher uncertainty ($p = 0.31, T = -1.00$, linear regression) or higher novelty ($p = 0.76$, $T = -0.29$, linear regression) (Fig. 2.1E). We then used a logistic regression to correlate decision variables and choices (see Materials and Methods section for details on logistic regression analysis), with EVs, uncertainty, and novelty as predictors. Model coefficients (Fig. 2.1F) indicated that patients were EV-seeking ($p < 10^{-4}$, $T = 5.15$, t-test) and novelty seeking ($p < 0.05$, $T = 2.26$, t-test), with a negative effect of the interaction between EV and trial number ($p < 0.01$, $T = -3.67$, t-test), suggesting a deviation from optimal outcome integration. Importantly, we also confirmed that patient behavior reflected value reset at the start of each block (see supplement), indicating that patients understood the task structure and learned how to choose the more advantageous options from their past experiences. Additionally, novelty and uncertainty correlated with behavior in a separable manner: while patients tended to be novelty-seeking overall, some patients avoided uncertainty whereas others were uncertainty seeking.

We compared two candidate computational models for explaining patient behavior (see Supplementary Material for details on model comparison). The first model (familiarity modulated upper confidence bound (fmUCB), Fig. 2.1G) was developed to describe behavior and fMRI data in a neurotypical population performing this same task (Cockburn et al., 2021). In this model, novelty promotes optimistic value initiation and modulates uncertainty to form an uncertainty bonus, which is added to q-values to construct stimulus utilities. The second model (linear novelty) relied on a linear combination between q-values, uncertainty and novelty to construct utilities. Using hierarchical Bayesian inference on patients' behavioral data (Piray et al., 2019) we determined the fmUCB model to explain behavior best, and a posterior predictive check showed that there are no systematic discrepancies between behavioral and simulated data. These behavioral modeling results show that a familiarity gating mechanism is appropriate to model the behavior of our participants. For subsequent neural data analyses we therefore used the variables derived from the fmUCB model as regressors.

PreSMA neurons represent stimulus features for individual options

We next probed the neural representation of stimulus features by examining whether the q-value, uncertainty, or novelty of each option presented on the screen was represented by neurons in our regions of interest using a Poisson GLM, with these features as regressors (for a complete list of encoding models, see Table 2.2). As these variables pertain to each stimulus being considered on a given trial and are not contingent on the choice of option that is subsequently made, they are candidate variables for acting as an input to the decision process.

We then grouped neurons according to their sensitivity to features associated with the left or right option, which we refer to as positional q-value, uncertainty bonus, and novelty neurons respectively (see Fig. 2.2D for an example). To determine whether activity in a brain area significantly correlated with these positional stimulus features, we tested whether the selected number of neurons were larger than expected by chance (Figs. 2.2A-C). All subsequent neuron count results were Bonferroni corrected for the number of time windows and brain areas in which we tested for a significant neuron count.

This analysis revealed prominent encoding of positional q-value, uncertainty bonus, and novelty during the trial onset period (19.9%, $p < 0.01/6$; 16.9%, $p < 0.05/6$; and 16.9%, $p < 0.05/6$, respectively, binomial test) and robust encoding of positional q-value during the pre-decision period in the preSMA (22.1%, $p < 0.001/6$, binomial test). In contrast, neurons in the vmPFC encoded positional q-value and uncertainty during the later pre-decision period, (17.4%, $p < 0.01/6$; 17.4%, $p < 0.01/6$, respectively), whereas novelty was encoded during both time periods (trial onset: 23.2%, $p << 10^{-5}/6$; pre-decision: 20.3%, $p < 0.001/6$, binomial test). None of the selected cell counts were significant in dACC (Figs. 2.2A-C). This indicates that preSMA and vmPFC neurons encode the variables which can serve as input to the decision process, aligned to trial onset in preSMA, and to decision in vmPFC.

Given the prominent role of preSMA and vmPFC in encoding all components of value in the trial onset period and the pre-decision period respectively, we investigated the temporal activity patterns for the selected neurons in these two areas. We repeated the Poisson GLM analysis described above in sliding time windows for visualization (Figs. 2.2E-F) and performed a Poisson latency analysis in neurons which were exclusively sensitive to one of the tested variables (Hanes, Thompson, and Schall, 1995) to compare onset latencies (see Materials and Methods).

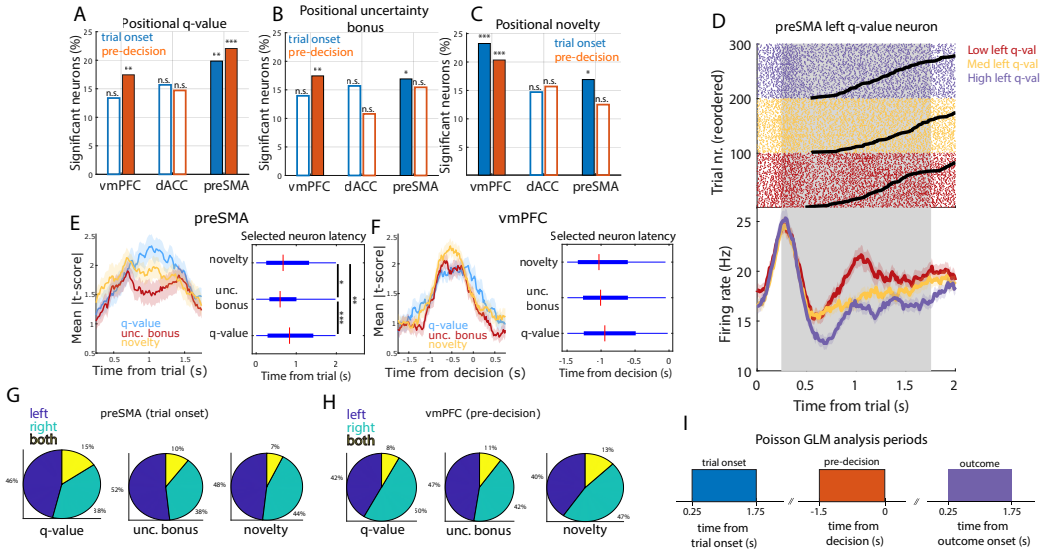

Figure 2.2: Encoding positional utility components in preSMA and vmPFC. (a) Percentage of neurons sensitive to positional q-value, in the trial onset (blue) and pre-decision (orange) periods (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$, binomial test). Hollow bars indicate non-significant counts. (b) Same, for positional uncertainty. (c) Same, for positional novelty. (d) Left q-value preSMA neuron. Top: Trial aligned spike raster plots. Black lines indicate the response time. For plotting, we sorted trials by q-value tertile (purple: high; yellow: medium; red: low). Bottom: Trial onset aligned PSTH (bin size = 0.2s, step size = 0.0625s). Shaded areas indicate standard error. (e) Selected preSMA neuron timing in the trial onset period. Left: Mean absolute t-score from the Poisson GLM analysis, in q-value (blue), uncertainty bonus (red), and novelty (yellow) neurons. Shaded areas indicate standard error. Right: Box plots of latency time across trials for all q-value, uncertainty bonus, or novelty neurons (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$, two-sided rank-sum test). (f) Same, for vmPFC neurons in the pre-decision period. (g) Proportion of preSMA sensitive neurons encoding positional q-values (left), uncertainty bonuses (center), or novelty (right), for one or both options. (h) Same, for vmPFC neurons in the pre-decision period. (i) Time windows for all analyses (trial onset, pre-decision, and outcome).

In preSMA (Fig. 2.2E), positional uncertainty neurons (median time: 0.59s) became active first, followed by positional novelty neurons (median time: 0.67s, $p < 0.05$, two-sided rank sum test) and positional q-value neurons (median time: 0.83s, $p < 10^{-6}$, two-sided rank sum test). Positional novelty neurons were found to be active significantly earlier than positional q-value neurons ($p < 0.01$, rank sum test). In vmPFC, median activation times relative to the time of decision were not significantly different for any of the selected sub-populations. Therefore, neurons in preSMA, but not in vmPFC, which are sensitive to distinct value components,

have significantly different latencies. Notably, preSMA q-value neurons have longer latencies than novelty and uncertainty bonus neurons.

Neurons coding for "positional" components of value were predominantly sensitive exclusively to a single option: for every variable, less than 15% of selected neurons in preSMA (trial onset) and in vmPFC (pre-decision) were found to be sensitive to features of both the left and right options available (Figs. 2.2G-H). This indicates that these neural sub-populations encode q-values, uncertainty, and novelty preferentially for individual stimuli instead of a total or averaged signal across options, which would be a possible interpretation if there had been prominent simultaneous encoding of both left and right options.

For completeness, we repeated the above analysis using values for q-value, uncertainty bonus and novelty as derived from the alternative linear novelty behavioral model. This revealed qualitatively similar results for positional q-value and novelty encoding. However, no significant neuron counts for positional uncertainty bonus encoding was found in any brain area. This highlights how the fmUCB model is specifically capable of describing uncertainty bonus representations accounting for neuronal activity.

Taken together, these findings suggest that stimulus features for the two available options are first encoded in the preSMA (trial onset period), followed by the vmPFC in the pre-decision period. This encoding was in the form of most neurons signaling stimulus features for one but not both options, which would be expected from a signal that serves as an input to the decision process.

PreSMA neurons encode an integrated utility signal for individual stimuli

To determine whether neurons represented an integrated utility for each decision option (incorporating EV, uncertainty and novelty), we used the utility signal derived from our fmUCB model. We performed a Poisson GLM encoding analysis (Figs. 2.3A,B) with left utility, right utility and decision as regressors. We found that a significant number of preSMA neurons encoded left utility after the trial onset (16.2%, $p < 10^{-5}/6$, binomial test), and in the pre-decision period (13.2%, $p < 0.001/6$, binomial test). Similarly, a significant number of preSMA neurons encoded right utility after trial onset (11.8%, $p < 0.01/6$, binomial test), and in the pre-decision period (11.8%, $p < 0.01/6$, binomial test). This result suggests that single neurons in the preSMA encode an integrated utility signal for individual choice options. Alternatively, it is possible that neurons correlating with utility in our

regression analysis are mostly reflecting the effects of q-value per se.

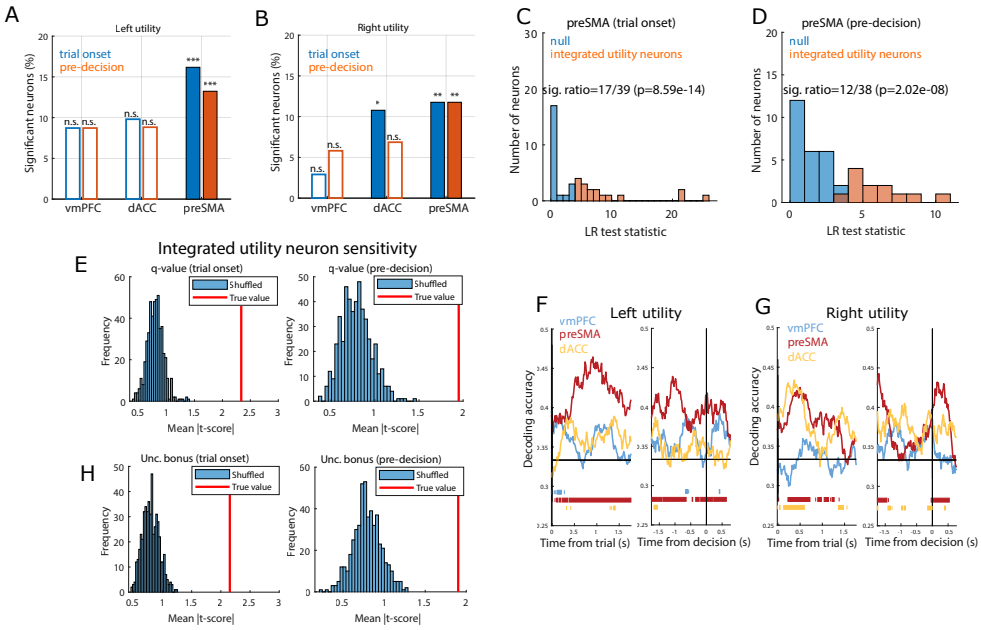

Figure 2.3: Neurons in preSMA encode integrated utility. (a) Percentage of left stimulus utility neurons in vmPFC, dACC, and preSMA, in the trial onset (blue) and pre-decision (orange) periods (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$, binomial test). Hollow bars indicate non-significant counts. (b) Same, for right stimulus utility. (c) Likelihood ratio test statistics across candidate preSMA integrated positional utility neurons, in the trial onset period. Neurons whose activity was better explained by a model containing q-values and uncertainty bonuses were classified as integrated utility neurons (orange). For the remaining ones (blue) the null model restricted to q-values was not rejected. (d) Same, for pre-decision period. (e) Integrated utility preSMA neuron sensitivity to q-values. Red lines indicate the mean absolute t-score across integrated utility neurons. Histograms include mean absolute t-scores for 500 iterations of bootstrapped null models with shuffled firing rates. Left: trial onset period; Right: pre-decision period. (f) dPCA population decoding performance for left utility for vmPFC (blue), preSMA (red), and dACC (yellow). Bars indicate periods of time where decoding accuracy was significantly above chance. Left: trial onset period; Right: pre-decision period. (g) Same, for right utility. (h) Same as (e), for uncertainty bonuses.

To test this hypothesis, we defined the sub-populations of preSMA neurons identified either as q-value or utility neurons as candidate neurons for an integrated utility signal. To determine whether they encoded an integrated utility signal versus q-value alone, we performed a likelihood ratio test ($p < 0.05$) comparing the performance of a model containing q-value, uncertainty, and decision regressors versus a null restricted model containing only q-value and decision (see Materials and Methods), while predicting each candidate neuron's spike count. The null restricted model

was rejected for 44% (17/39) of preSMA candidate neurons at trial onset and for 32% (12/38) of preSMA candidate neurons in the pre-decision window (Figs. 2.3C-D). Therefore, a significant portion of candidate neurons in preSMA qualified as integrated utility neurons (trial onset: $p < 10^{-13}$; pre-decision: $p < 10^{-7}$). These integrated utility neurons collectively encoded the main components of utility (q-values and uncertainty bonuses) at a higher level than expected by chance ($p < 0.002$ in all instances, permutation test), further confirming their role in computing an integrated signal (Figs. 2.3E,H).

Population decoding of integrated utility as an input to the decision process

Building upon these results demonstrating that single neurons in preSMA and vmPFC encode stimulus features that could support the decision making process, we next tested when and where it was possible to decode an integrated stimulus utility value from neural populations. To do so, we consider the firing patterns of all neurons from each brain region across all trials, employing demixed principal component analysis(Kobak et al., 2016) (dPCA) to reduce the data dimensionality (see Materials and Methods).

We performed two separate analyses for left and right utilities, including the decision itself (i.e. left vs right choice) as a marginalization in both analyses (Figs. 2.3F-G). Left and right option utility was decodable in preSMA, both following trial onset and preceding the button press (Figs. 2.3F-G, significant time periods are indicated in the figure). Compatible with the cell selection results, neither left nor right side utility was robustly decodable from vmPFC.

Thus, these results suggest that preSMA encodes an integrated utility signal that encompasses both q-values and uncertainty. At the population level, the utility for each decision option was decodable in preSMA even after demixing utility from the decision, indicating that the utility for each of the two possible decision options is represented at the population level. Together, these findings suggest that preSMA neurons represent the signals needed as an input to the decision making process.

Decision is represented later than stimulus utility

At the level of single neurons, the decision was encoded in the preSMA only in the pre-decision period (Fig. 2.4A), in which 14.0% ($p < 10^{-4}/6$, binomial test) of neurons were decision selective (Fig. 2.4F shows an example). In preSMA, neurons that encoded left or right utility were largely distinct from those encoding the decision (Fig. 2.4C), with no significant overlap for either ($p = 0.22$ and

p = 0.28, Jaccard test, see Materials and Methods). Neither at the single-neuron (Fig. 2.4A) nor the population level was the decision represented in vmPFC or dACC (Fig. 2.4D), indicating a privileged role for the preSMA in representing choices in our task. We therefore restrict the following analysis to the preSMA.

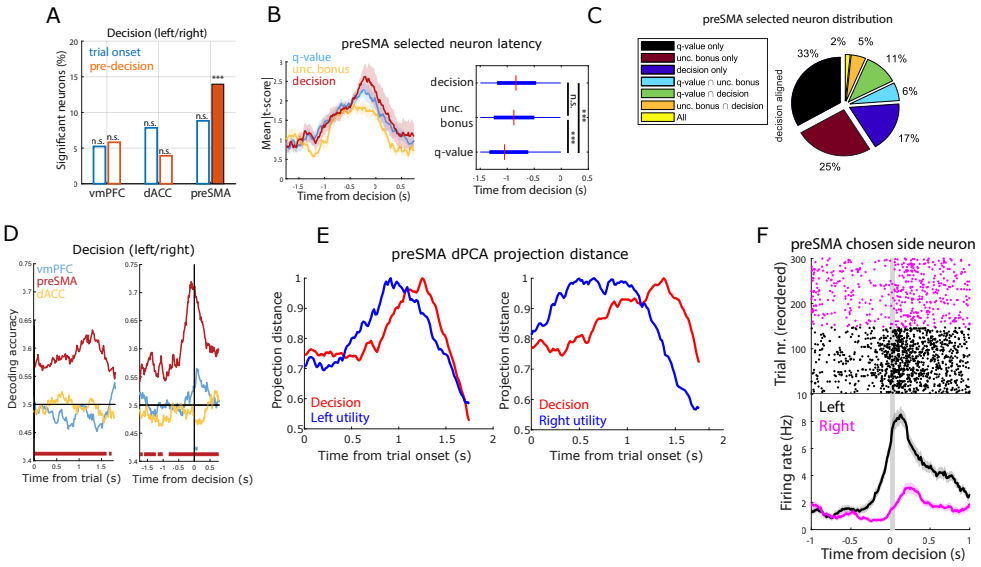

Figure 2.4: PreSMA encodes decisions. (a) Percentage of decision neurons (left vs. right choice) in vmPFC, dACC, and preSMA, in the trial onset (blue) and pre-decision (orange) periods (*** = $p < 0.001$, binomial test). Hollow bars indicate non-significant counts. (b) Sensitive preSMA neuron timing in the pre-decision period. Left: Mean absolute t-score for q-value (blue), uncertainty bonus (yellow), and decision (red). Shaded areas indicate standard error. Right: Latency time box plots for all q-value, uncertainty bonus, or decision neurons (*** = $p < 0.001$, two-sided rank-sum test). (c) Proportion of preSMA neurons (pre-decision period) for q-value, uncertainty bonus, decision and combinations thereof. (d) dPCA decision decoding for vmPFC (blue), preSMA (red), and dACC (yellow). Bars indicate significant times, comparing to a bootstrapped null distribution. Left: trial onset period; Right: pre-decision period. (e) Normalized Euclidean distance between dPCA projections onto principal utility components (blue), between low and high utility trials, and decision components (red), between left and right decision trials, with left (left) or right (right) utility marginalizations. (f) Example preSMA decision neuron. Top: Raster plot. For plotting, we sorted trials in left (black) and right (magenta) decisions. Bottom: PSTH (bin size = 0.2 s, step size = 0.0625 s). Gray bar indicates button press. Shaded areas indicate standard error.

Relative to the time of response, a single-unit analysis showed that q-value neurons responded first at $-1.04s$, earlier than uncertainty bonus neurons at $-0.87s$ ($p < 10^{-3}$, two-sided rank-sum test), and decision neurons at $-0.83s$ ($p < 10^{-3}$, two-sided rank-sum test). At the population level, we projected neural data onto the dPCA

demixed principal components components separately for low/high utility trials, and for left/right decisions. We then examined the Euclidean distances between these trajectories as a function of time. This revealed that the distance in state space was maximal for positional utility earlier than for decisions (Fig. 2.4E). Relative to trial onset, this latency difference was apparent for both left utility (0.91s vs. 1.25s) and right utility (0.65s vs. 1.37s).

Therefore, in the preSMA, decisions and stimulus values are encoded by largely separate groups of neurons, with utility encoding appearing earlier than the decision. This time course and encoding scheme suggests that preSMA encodes pertinent stimulus features pre-decision, thereby revealing a potential substrate for value-based decision-making.

Decision conditioned variables are represented in vmPFC and dACC neurons

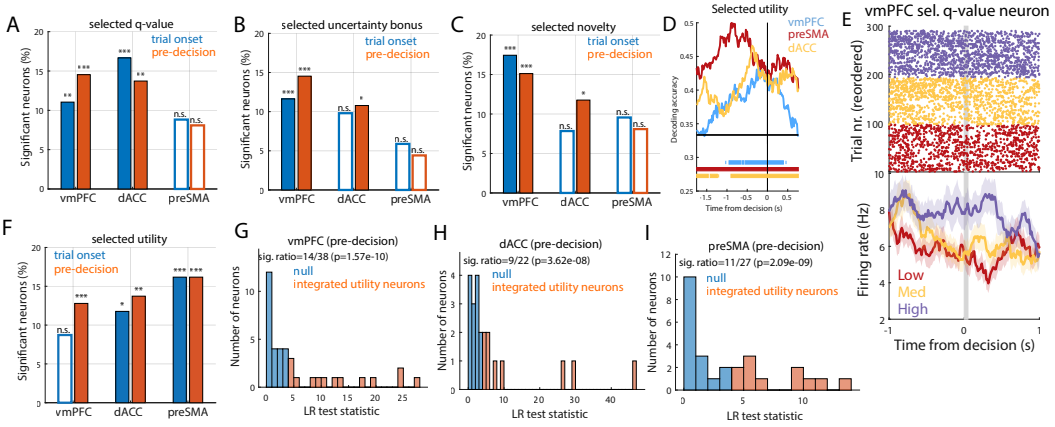

Figure 2.5: Encoding selected stimulus properties. (a) Percentage of selected q-value neurons in vmPFC, dACC, and preSMA, in the trial onset (blue) and pre-decision (orange) periods (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$, binomial test). Hollow bars indicate non-sensitive counts. (b) Same, for selected uncertainty. (c) Same, for selected novelty. (d) dPCA selected utility decoding in the pre-decision period, for vmPFC (blue), preSMA (red), and dACC (yellow). Bars indicate significant decoding accuracies for each brain region, comparing to a bootstrapped null distribution. (e) VmPFC selected q-value neuron. Top: Raster plots. For plotting, we sorted trials by q-value tertiles (purple: high; yellow: medium; red: low). Bottom: PSTH (bin size = 0.2 s, step size = 0.0625 s). Gray bar indicates button press. Shaded areas indicate standard error. (f) Same as (a), for selected utility. (g) Histogram of likelihood ratio test statistics across candidate vmPFC integrated selected utility neurons (orange), in the pre-decision period. For the remaining neurons (blue) a null model containing only selected q-values was not rejected. (h) Same as (g), for dACC. (i) Same as (g), for preSMA.

Representing the expected outcome of a choice is a critical step in decision making as it facilitates learning by way of comparison to observing the actual outcome received. We therefore next examine the neuronal representation of selected option's utility (see full selection-based model in Table 2.2). Components of the selected option's utility were encoded in vmPFC and dACC, but not in preSMA. In vmPFC, selected q-values, uncertainty, and novelty were encoded in both the trial onset and pre-decision period (Fig. 2.5A-C, Fig. 2.5E shows an example). In dACC, all three variables were also encoded in the pre-decision period. We also found that selected novelty neurons became active significantly earlier in vmPFC than dACC (-1.06s vs. dACC: -0.80s, $p < 0.05$), with no significant latency differences between the two areas for the other two variables (Fig. 2.11A). Furthermore, we examined encoding of value for the rejected option (see Supplementary Material, Fig. 2.8A-C). Additionally, a significant proportion of vmPFC selected uncertainty neurons signal whether a trial is exploratory or not prior to button press (see Supplementary Material). Overall, these findings indicate that single neurons in vmPFC and dACC encode value components contingent on the decision that was made.

Similar to the integrated positional utility analysis, we defined a group of candidate integrated selected utility neurons as the subset of units that correlated with the selected option's q-value or utility (shown in Fig. 2.5F). We determined that, in all brain areas, the number of neurons selected this way was larger than expected by chance (Figs. 2.5G-I) (vmPFC: $p < 10^{-9}$; dACC: $p < 10^{-7}$; preSMA: $p < 10^{-8}$). The activity of this subset of neurons was therefore indicative of an integrated selected utility.

Finally, we examined the points in time at which the selected option's integrated utility could be decoded from pooled activity across all neurons in the regions of interest (using dPCA, see Materials and Methods). This analysis revealed robust decoding of selected utility in all brain areas, with a notably earlier onset in preSMA compared to both vmPFC and dACC (Fig. 2.5D). Motivated by the earlier utility decoding in preSMA, we tested whether selected utility neuron latency times were also shorter in preSMA than in the other areas. A Poisson latency analysis revealed an onset time in preSMA of $-0.83s \pm 0.01$, which was significantly earlier than in vmPFC ($-0.79s \pm 0.01$, $p < 0.05$, one-sided rank sum test) and dACC ($-0.71s \pm 0.02$, $p < 10^{-5}$, one-sided rank sum test).

Taken together, these findings establish widespread value coding specific to the chosen option in all tested brain areas. One interpretation of these findings is that

features of selected stimuli are monitored after the decision in the time window that immediately precedes the button press. While all areas displayed evidence of integrated selected utility coding, the preSMA represented this signal earlier than the other regions, consistent with the possibility that the preSMA is more closely involved in the choice process.

Post-feedback neuronal responses

Behavioral consequences offer information that can be leveraged to make better decisions in the future. We tested for neurons encoding reward information, probing for representations of outcome, expected value and RPE during the feedback period. (Figs. 2.6A-D). Outcome (win/lose) was robustly encoded in dACC, preSMA and vmPFC (Percentage of neurons selected 34.3%, $p = 0$; 35.3%, $p = 0$; and 17.4%, $p < 10^{-9}/3$, respectively, binomial test). The q-value of the selected stimulus was encoded in vmPFC and preSMA, but not dACC (12.2%, $p < 10^{-4}/3$ and 15.4%, $p < 10^{-5}/3$, respectively).

We probed for two forms of the RPEs: the RPEs absolute value tracking surprise irrespective of valence, and a proxy for signed RPE, outcome minus selected q-value. Signed RPE was encoded in vmPFC and preSMA (11.6%, $p < 10^{-3}/3$ and 16.9%, $p < 10^{-7}/3$, respectively), but not dACC (Figs. 2.6D). In contrast, we did not find significant numbers of neurons encoding the RPEs absolute value in either brain area (Figs. 2.6C).

Latency analysis revealed that contrary to error signals we have studied previously (Fu et al., 2019), dACC neurons encoded outcome significantly earlier than both preSMA and vmPFC (Fig. 2.6I; outcome-aligned median latency: 0.50s vs. 0.79s, $p < 10^{-17}$ and vs. 0.81s, $p < 10^{-7}$, two-sided rank-sum test). There was no difference between the onset of outcome signals in preSMA and vmPFC ($p = 0.21$, rank-sum test). Lastly, outcome neurons became active earlier than selected q-value neurons in all three regions (median times: vmPFC: 0.91s, $p < 0.05$; dACC: 0.69s, $p < 0.001$; preSMA: 0.97s, $p < 10^{-4}$, rank-sum test), indicating that selected q-value representations were not persistently maintained from the choice period.

Whereas the majority of neurons encoded one of these three variables exclusively, approximately one fourth encoded multiple variables (Figs. 2.6K; proportion of mixed neurons out of all sensitive neurons: vmPFC: 19%; dACC: 16%; preSMA: 25%). We cautiously speculate that this mixed selectivity could be accounted for by independent probabilities of a neuron representing a given variable. In preSMA,

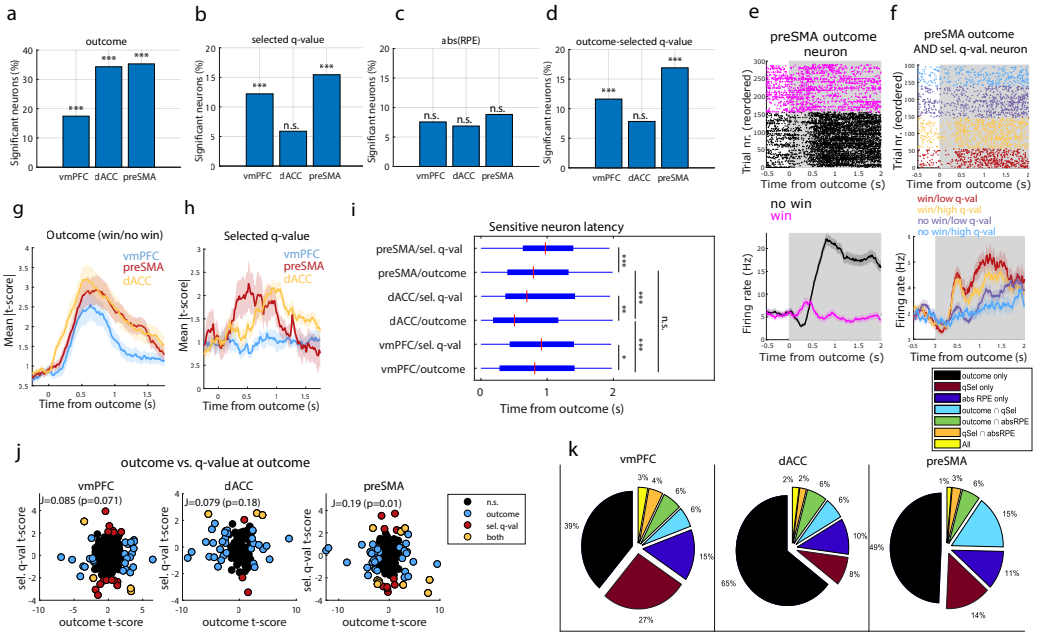

Figure 2.6: Post-feedback encoding. (a) Percentage of outcome neurons in vmPFC, dACC, and preSMA (*** = p<0.001, binomial test). (b) Same, for selected q-value. (c) Same, for abs(RPE). (d) Same, for the outcome minus selected q-value contrast. (e) Outcome neuron in preSMA. Top: Raster plots. For plotting, we sorted trials by outcome (magenta: win; black: no-win). Bottom: PSTH (bin size = 0.2 s, step size = 0.0625 s). Shaded areas indicate standard error. (f) Same, for an outcome and selected q-value preSMA neuron. Trials were split into outcome/q-value groups: win/low (red); win/high (yellow); no-win/low (purple); no-win/high (blue). (g) Mean absolute t-score in outcome neurons in vmPFC (blue), dACC (red), and preSMA (yellow). (h) Same, for selected q-value neurons. (i) Latency times box plot for outcome or selected q-value neurons in vmPFC, dACC, or preSMA. (* = p<0.05; ** = p<0.01; *** = p<0.001, two-sided rank-sum test). (j) Scatter plot of outcome versus selected q-value t-scores. We display neurons sensitive to outcome (blue), to selected q-value (red), or both (yellow). We indicate the Jaccard overlap index and a p-value for the Jaccard test. Left: vmPFC; Center: dACC; Right: preSMA. (k) Pie charts of neuron preference for outcomes, selected q-values, abs(RPE). Left: vmPFC; Center: dACC; Right: preSMA.

the proportion of neurons that signaled both outcome and selected q-value was higher than expected by independence ($p < 0.05$, Jaccard index test), but not in vmPFC ($p = 0.07$, Jaccard index test) or dACC ($p = 0.18$, Jaccard index test). This suggests that the activity of individual neurons in preSMA uniquely contains sufficient information to support the computation of reward prediction errors.

2.4 Discussion

We investigated value-based decision making in human single neurons, while manipulating variables relevant to resolution of the explore/exploit dilemma; specifically, stimulus value, uncertainty and novelty. By recording from three areas of the prefrontal cortex implicated in decision-making across humans and other animals(Wunderlich, Rangel, and O'Doherty, 2009; Goñi et al., 2011; Hare, Schultz, et al., 2011; Grabenhorst and Rolls, 2011; M. F. Rushworth et al., 2012; Strait, Blanchard, and Hayden, 2014; Li et al., 2016; Hunt et al., 2018; Domenech, Rheims, and Koechlin, 2020; B. Averbeck and O'Doherty, 2021), we identified how these variables are encoded, and addressed how they are integrated to inform decisions. Our findings highlight a particularly important role for human preSMA neurons in value-based decisions.

We found evidence for separate representations of the EV, uncertainty and novelty associated with options under consideration in human single neurons in both the preSMA and vmPFC, supporting the separable encoding of each of these decision variables across these areas. Crucially, we also found that a subset of EV coding neurons were better explained by an integrated utility signal, in which the option's EV was combined with uncertainty and novelty. This signal was most robustly represented in the preSMA, where it was encoded both at the single neuron and population levels. These findings provide a proof of principle for the existence of an integrated utility signal in human prefrontal neurons.

We also identified a distinct population of preSMA neurons encoding the decision itself above and beyond stimulus utility, expanding on previous findings linking preSMA to volitional decision making (Fried, Mukamel, and Kreiman, 2011). Thus, unlike dACC or vmPFC, the preSMA represented not only the key utility signal which informs choice, but also the behavioral product of the decision itself. These results for value-based decisions expand on previous work which reported choice signaling in categorization and memory tasks in preSMA and dACC (Minxha, Adolphs, et al., 2020). We found robust outcome tracking in dACC and preSMA, in consonance with previous findings in human dmPFC (Gazit et al., 2020), and while preSMA (as well as SMA(Bonini et al., 2014)) had been shown to monitor internally generated error responses, preceding dACC error neurons temporally (Fu et al., 2019), we observed that value-based feedback elicited earlier outcome responses in dACC than in preSMA. Reward prediction error, on the other hand, was more robustly encoded in preSMA than in dACC. Taken together, these results

appear to position the preSMA as playing a central role in value-based decision-making in humans, particularly in decision tasks that elicit the integration of multiple stimulus features as is required to balance the explore/exploit trade-off. Although we found that the preSMA plays a privileged role in encoding decision variables, we expect that these computations are likely supported by a broader cortico-striatal network beyond the preSMA alone (Kim and Shadlen, 1999; Nambu, Tokuno, and Takada, 2002; Haber and Knutson, 2010; Ding and Gold, 2010; Yartsev et al., 2018; Fan, Gold, and Ding, 2020; Chen et al., 2020).

Our findings support a distinction between dorsal and ventral areas of cortex, whereby dorsal regions contribute to action-based decisions while more ventral areas such as the vmPFC are involved in valuation but not in decisions over actions (Walton, Behrens, et al., 2010; Noonan, Mars, and M. Rushworth, 2011; Rudebeck and Murray, 2011; M. F. Rushworth et al., 2012; Domenech and Koechlin, 2015; Murray and Rudebeck, 2018; Domenech, Rheims, and Koechlin, 2020). Here we find that a similar organization applies at the level of human single neurons. However, we also found a degree of specificity within the dorsal human prefrontal cortex as to where integrated utility and the decision itself are encoded: in preSMA but not as robustly in dACC. These findings situate the human preSMA as being more prominently involved in the computations directly required for value-based decision-making than the sub-region of dACC from which we recorded. The present findings thus contribute to a more fine-grained understanding of functional specificity within dorsomedial prefrontal cortex.

We also looked for the representation of variables pertinent to the selected option; and thus, contingent on the decision made. The integrated utility for the option that was ultimately chosen was found to be widely encoded throughout all three of the brain regions we recorded. It is noteworthy that this signal emerged markedly earlier in the preSMA than in vmPFC, consistent with the possibility that preSMA is more proximal to the generation of the decision itself than is the vmPFC. Unlike the preSMA, single neuron activity in vmPFC also correlated with individual decision variables for the value, uncertainty and novelty of those stimuli that had been selected on a given trial. Furthermore, a significant portion of vmPFC neurons encoding selected uncertainty were also modulated by whether a decision was classified as exploratory or not (Fig. 2.9). When taken together, these findings suggest a role for vmPFC neurons in post-decisional monitoring of option features, especially in the context of exploratory decision making.

We found widespread outcome encoding across all three regions, in consonance with a vast literature implicating prefrontal cortex in signaling outcomes, in rodents (Pratt and Mizumori, 2001; Gutierrez et al., 2006; Horst and Laubach, 2013; Malvaez et al., 2019), monkeys (Amiez, Joseph, and Procyk, 2006; M. Matsumoto, K. Matsumoto, Abe, et al., 2007; Kennerley, Behrens, and J. D. Wallis, 2011; Knudsen and J. D. Wallis, 2020), and humans (Li et al., 2016; Hill, Boorman, and Fried, 2016). We further found significant evidence for concurrent encoding of outcomes and selected EVs in preSMA post-feedback, which together constitute the two main components of reward-prediction errors (R. Rescorla and Wagner, 1972; Sutton, 1988; Sutton and Barto, 2018). These findings suggest that preSMA neurons can support learning of reward expectations.

In conclusion, our results situate the human preSMA as an important center for value-based decision-making, with a robust encoding of decision variables and, most crucially, an integrated utility signal at the single neuron level that can be leveraged to inform choice. While vmPFC neurons encoded pre-decision variables as well as post-decision variables contingent on choice, neither this region nor the dACC showed an equivalently robust encoding of pre-decision integrated utilities or the choice itself. These findings suggest that value-based decision-making during exploration depends on highly specialized computations performed in distinct areas of the prefrontal cortex. Furthermore, the existence of an integrated utility at the level of single neurons that could serve as the input to the choice process suggests that relevant decision variables such as EV, uncertainty, and novelty are first integrated into a unified neuronal representation prior to being entered into a decision comparison, shedding light on how subjective utility-based choices are implemented in the human brain.

2.5 Materials and Methods

Electrophysiology and recording

We used Behnke-Fried hybrid depth electrodes (AdTech Medical), positioned exclusively according to clinical criteria (Supplementary Table 2.3). Broadband extracellular recordings were performed with a sampling rate of 32 kHz and a bandpass of 0.1-9000Hz (ATLAS system, Neuralynx Inc.). The data set reported here was obtained bilaterally from ventromedial prefrontal cortex (vmPFC), dorsal anterior cingulate cortex (dACC), and pre-supplementary motor area (preSMA) with one macroelectrode on each side. Each macroelectrode contained eight 40 μm microelectrodes. Recordings were bipolar, utilizing one microelectrode in each bundle of

eight microelectrodes as a local reference.

Patients

Twenty patients (fourteen females) were implanted with depth electrodes for seizure monitoring prior to potential surgery for treatment of drug resistant epilepsy. Two of the patients performed the task twice, totalling 22 recorded sessions. Human research experimental protocols were approved by the Institutional Review Boards of the California Institute of Technology and the Cedars-Sinai Medical Center. Electrode location was determined based on preoperative and postoperative T1 scans obtained for each patient.

Task

Patients performed a two-armed bandit task (Fig. 2.1B). The task contained 20 blocks of 15 trials, for a total of 300 trials. The 20 blocks were split into 2 recording sessions with 10 blocks each, with a 5 minute break in between sessions. Each trial began with a baseline period (sampled randomly from a uniform distribution of 0.75-1.25s), followed by a choice screen showing the two available slot machines presented on the left or on the right of the screen. The identity of each slot machine was uniquely identifiable by a painting displayed on the center of each slot machine. Patients had to decide between the left or the right option by pressing a button in less than 3s, or the trial would be considered missed and no reward would be accrued. Across all trials, mean reaction time (RT) was $1.47s \pm 0.02$ (relative to onset of choice screen). Following the button press, the chosen slot machine is shown for a period of 1-2s (sampled randomly from a uniform distribution), followed by the outcome screen shown for 2s. The outcome screen showed either a golden coin to represent winning a reward, or a crossed-out coin to represent not winning (both shown on top of the chosen slot machine).

To shape the novelty and uncertainty of presented stimuli, we manipulated which stimuli would appear in each block and each trial according to the rules described as follows. For each block, the identity of the two slot machines that appeared in each trial was drawn randomly from a set of 3 possible options, selected specifically for each block. In the first block, the 3 options were selected randomly from a set of 200 paintings. In every subsequent block, one out of the three stimuli from the previous block was chosen to be replaced, substituting it for a novel unused stimulus out of the 200 paintings.

To manipulate the interaction between stimulus novelty and trial horizon, in every

block after the first one, we chose stimuli to be held out and only presented after a minimum trial threshold, selected randomly for each block, between 7 and 15 trials. For every block after the first one, we alternated whether the held out stimulus would be one of the familiar ones or the novel stimulus for that block.

The probabilities of receiving a reward from each slot machine were reset in the beginning of every block, and determined according to the chosen difficulty of each block, which alternated between the easy and hard conditions. Crucially, these reward probabilities did not change within each block. In the easy condition, reward probabilities were more widely spaced out between different slot machines, and chosen from the values $[0.2, 0.5, 0.8]$. In the hard condition, the possible probabilities were $[0.2, 0.5, 0.6]$.

Some patients performed a shorter variant of the task, which consisted of 206 trials across 10 blocks (see Supplementary Table 2.1). In this version, the set of possible stimuli in each block contained 5 options, in each block after the first one 2 novel options were introduced, one of which composed the held out set along with one out of the 3 familiar options from the previous blocks. Bandit win probabilities were sampled from the linearly spaced interval $[0.2, 0.8]$ in easy blocks and from $[0.4, 0.6]$ in hard blocks. We pooled data from the two task variants together for all analysis.

Behavioral analysis and computational modeling

Logistic regression for value components and decisions

We used a logistic regression model to describe how the past history of rewards, sampling history, stimulus exposure history, and their interactions with trial number correlated with decisions (Fig. 2.1F). For this, we defined q-value Q_s as the mean of a beta distribution which estimates the probability of receiving a reward from a bandit, as determined by the history of wins and losses after sampling a stimulus s, as well as $\delta Q = Q_{left} - Q_{right}$, the difference between left and right Q-values. Similarly, we define an uncertainty value U as the variance of the same beta distribution, as well as its corresponding differential $\delta U = U_{left} - U_{right}$. Finally, we defined novelty N as the variance of a beta distribution in which $\beta = 1$ and the α parameter is the number of times patients were exposed to a stimulus s in the entire session, as well as its corresponding differential $\delta N = N_{left} - N_{right}$:

We then performed a logistic regression using MATLAB's function $mnrfit$ to model the probability p_{left} of a left decision based on these regressors as well as

their interaction with the trial number t within a block:

$$\log \frac{p_{left}}{1 - p_{left}} = \beta_0 + \beta_1 \delta Q + \beta_2 \delta U + \beta_3 \delta N + \beta_4 \delta Q \cdot t + \beta_5 \delta U \cdot t + \beta_6 \delta N \cdot t \quad (2.1)$$

Familiarity gating model of exploration (fmUCB)

We compared two computational models fit to patients' behavior. Individualized model fits and model comparisons were obtained across the patient population through hierarchical Bayesian inference (Piray et al., 2019). This method yielded model parameters for each subject in the data set, for each of the tested models, as well as exceedance probabilities, which expressed the probability that either model was the most frequent in the behavioral dataset (Rigoux et al., 2014).

The first model we tested is a fmUCB model (Cockburn et al., 2021) of exploratory decision making. In this model, the choice probability for a decision d in a trial t is estimated using the utilities assigned to the left (U_L) and right (U_R) options, through a softmax function:

$$p_t(d = LEFT) = \frac{1}{1 + e^{\beta(U_{R,t} - U_{L,t})}} \quad (2.2)$$

In this equation, β is the inverse temperature free parameter. To balance incentives to explore and exploit different stimuli, the utilities assigned to each stimulus s on a trial t were defined to be the sum of its weighted q-values and an uncertainty bonus B, depending on the past history of rewards received from the slot machine, and how often the slot machine had been sampled, respectively:

$$U_{s,t} = Q_{s,t} + B_{s,t} \quad (2.3)$$

The q-value was defined similarly to the expected value of a beta distribution, as a function of the past history of wins and losses received from a slot machine, modified to account for the effect of recency over stimulus preferences:

$$Q_{s,t} = \frac{\alpha_{s,t}}{\alpha_{s,t} + \beta_{s,t}} \quad (2.4)$$

In this equation, $\alpha_{s,t}$ and $\beta_{s,t}$ describe the effect of previous wins and previous losses, respectively, received from the slot machine s before trial t.

The α term is defined as follows, where $H^W_{s,t}$ is how many times sampling slot machine s has resulted in a win before trial t, and w is an exponentially decaying effect of recency. The time scale of this exponential decay is determined by a learning rate free parameter λ, fit in the interval (0,1):

$$\alpha_{s,t} = 1 + \sum_{i=1}^{t-1} w_{i,t} H^W_{s,t} \tag{2.5}$$

$$w_{i,t} = (1 - \lambda)^{(t-i)} \tag{2.6}$$

Similarly, the β term is defined as follows, where $H^L_{s,t}$ is how many times sampling slot machine s has resulted in a no-win before trial t:

$$\beta_{s,t} = 1 + \sum_{i=1}^{t-1} w_{i,t} H^L_{s,t} \tag{2.7}$$

We also allowed novelty to bias the initialization of the α and β hyperparameters, to include an optimistic initialization strategy (Wittmann et al., 2008) for exploration. This was done by including a novelty initialization bias free parameter n_I, which was modulated by the same exponential decay $w_{0,t}$, creating the novelty bias $n_I w_{0,t}$. If $n_I w_{0,t} > 0$, we would add this quantity to $\alpha_{s,t}$, resulting in a novelty seeking bias, and if $n_I w_{0,t} < 0$, we added this quantity to $\beta_{s,t}$, resulting in a novelty avoidance bias.

The uncertainty bonus term in Equation 2.3 was defined as a function of raw stimulus uncertainty, gated by familiarity, and weighed by each patients' uncertainty preferences, according to the weight parameter w^U_t, as a function of the trial horizon within a block, as will be defined further:

$$B_{s,t} = V_{s,t} F_{s,t} w^U_t \tag{2.8}$$

Raw stimulus uncertainty $V_{s,t}$ was defined similarly to the variance of a beta distribution, as a function of how many times a stimulus has been sampled, using the previously defined $\alpha_{s,t}$ and $\beta_{s,t}$ terms:

$$V_{s,t} = 12 \frac{\alpha_{s,t} \beta_{s,t}}{(\alpha_{s,t} + \beta_{s,t})^2 (\alpha_{s,t} + \beta_{s,t} + 1)} \tag{2.9}$$

We introduced a normalizing factor of 12 to the raw stimulus uncertainty equation to ensure that maximal uncertainty, obtained when $\alpha_{s,t} + \beta_{s,t} = 1$, is equal to 1.

The familiarity gating was introduced to allow for the novelty of a stimulus, as a function of how many times it has been seen throughout the session, to interact with the behavioral effects of uncertainty. Defining g as a familiarity gating free parameter, fit for each subject, the familiarity gating $F_{s,t}$ is defined as follows:

$$F_{s,t} = 1 - gN_{s,t} \tag{2.10}$$

In this equation, $N_{s,t}$ is the novelty value for slot machine s in trial t, defined as a monotonically decreasing function of the number of exposures for s, defined as $E_{s,t}$:

$$N_{s,t} = 12\frac{E_{s,t} + 1}{(E_{s,t} + 2)^2(E_{s,t} + 3)} \tag{2.11}$$

This definition of novelty was chosen to create a similar functional form to the uncertainty value, while enforcing that maximal novelty, for stimuli that had not been exposed before, was equal to 1.

Finally, to allow for switching between exploration and exploitation within a block, the effect of trial horizon over the uncertainty bonus was defined a linear function of the trial number within a block, adding the free parameters for terminal uncertainty uT and uncertainty intercept uI, where T is the maximal trial horizon for a block and uS is the uncertainty slope:

$$uS = \frac{(uT - uI)}{T} \tag{2.12}$$

$$w_t^U = uI + uS(t - 1) \tag{2.13}$$

Alternative model with independent utility of novelty (linear novelty model)

In the second model we tested, uncertainty and novelty did not interact directly in the construction of the uncertainty bonus, but are added as independent components of the stimulus utility value. Concretely, Equation 2.3 is modified to add a novelty bonus $N^*_{s,t}$ to utility:

$$U_{s,t} = Q_{s,t} + B^*_{s,t} + N^*_{s,t} \tag{2.14}$$

The uncertainty bonus from Equation 2.8 was adapted, creating a modified uncertainty bonus $B^*_{s,t}$, to remove the interaction with novelty through familiarity gating:

$$B^*_{s,t} = V_{s,t} w^U_t \tag{2.15}$$

We defined the novelty bonus $N^*_{s,t}$ similarly to the uncertainty bonus, by multiplying the previously defined novelty value (Equation 2.11) by a novelty weight free parameter (w^N_t), which was fit for each patient:

$$N^*_{s,t} = N_{s,t} w^N_t \tag{2.16}$$

The remaining components of the alternative model with a novelty bonus are kept the same as in the fmUCB model.

Neural data pre-processing

We performed spike detection and sorting with the semiautomatic template-matching algorithm OSort (Rutishauser, Schuman, and Mamelak, 2006). Channels with interictal epileptic activity were excluded. Across all 22 sessions, we obtained 191 vmPFC, 137 preSMA and 108 dACC putative single units (436 total). In this manuscript we refer to these isolated putative single units as "neuron" and "cell" interchangeably. For the single neuron encoding analyses in this study we pre-selected only neurons with more than 0.5Hz average firing rate across all trials, resulting in 172 vmPFC, 136 preSMA and 102 dACC putative single units (410 total).

Poisson GLM encoding analysis

We used Poisson regression GLMs to select for neurons, with response variable the number of spikes fired and the dependent variable different subsets of model variables. We computed the spike counts in every trial in four windows of interest (trial onset, from 0.25s to 1.75s, aligned to trial onset; pre-decision, from -1s to 0s, aligned to button press; and outcome, from 0.25 to 1.75s, aligned to outcome onset). For visualization purposes, we also fit the same models with 0.5s time windows, sliding by 16ms steps, within the same time limits. We then tested hypotheses about how the spike count of each neuron was correlated with left and right utility (U_L,U_R), chosen side ($Side$), left and right q-value (Q_L,Q_R), left and right uncertainty bonus (B_L,B_R), left and right novelty (N_L,N_R), as well as their selected and rejected counterparts, outcome (O), and absolute reward prediction

error ($|RPE|$). Additionally, we performed these encoding analyses utilizing the raw uncertainty values $V_{s,t}$ instead of the transformed uncertainty bonus values $B_{s,t}$, and obtained equivalent results.

We also tested whether neuronal activity in the pre-decision period correlated with whether a trial was classified as an explore or a non-explore trial, correcting for selected uncertainty bonus. We defined explore trials as those in which $Q_{sel} < Q_{rej}$ and $U_{sel} > U_{rej}$, defining the explore flag $Explore = 1$ for those trials and $Explore = 0$ for all others. For these analyses, we specified the models described on Supplementary Table 2.2 and fit them with the MATLAB function $fitglm$.

To understand the overall role of q-values, uncertainty and novelty regardless of the position of stimuli on the screen (Fig. 2.2), we fit the full positional model coefficients and reported the proportion of sensitive neurons for the left and right components together, as positional q-value neurons, positional uncertainty bonus neurons and positional novelty neurons.

For the outcome analysis (Fig. 2.6), we performed an F-test for the difference between coefficients for outcome and selected q-value ($b_1 - b_2$), as a proxy for reward prediction error coding, and reported the proportion of neurons for which the contrast is different than 0. We also report the number of neurons whose activity correlates with outcomes and absolute reward prediction error at the time of outcome.

Poisson latency analysis

To determine when individual neurons became active at a single trial level, we performed Poisson latency analyses(Hanes, Thompson, and Schall, 1995) for pre-selected groups of neurons sensitive to the variable of interest in the encoding analyses (Figs. 2.2 E-F; Fig. 2.4B; Fig. 2.6I). This method detects the first point in time in which interspike intervals significantly differ from what would be expected from a constant firing rate Poisson point process, using the neuron's average firing rate as the rate parameter. We used a significance parameter of $p < 0.05$ as our burst detection threshold for all analyses.

Jaccard index test

After performing Poisson GLM encoding analyses, we tested whether the sub-populations of neurons which were sensitive to two variables of interest had significant overlap. For this, we computed the Jaccard index (Jaccard, 1912) of overlap

between neurons sensitive to each of the variables X and Y, where N_X and N_Y indicate the number of neurons sensitive to the variables X and Y, respectively, and $N_{X,Y}$ indicates the number of neurons concurrently sensitive to both variables:

$$J = \frac{N_{X,Y}}{N_X + N_Y - N_{X,Y}} \tag{2.17}$$

To compute p-values for each comparison between two variables, we bootstrapped a null distribution of Jaccard indexes using 1000 reshuffles, considering X and Y are independent variables with a false positive rate of $p = 0.05$.

Likelihood ratio hypothesis testing

We tested whether neurons in positional q-value or positional utility sensitive sub-populations had their activity better explained by an unrestricted model including the main additive components of utility (q-value and uncertainty bonus) or by a restricted model including only q-values, given the correlations we observed between q-values and integrated utility values. Neurons which had their activity better explained by the unrestricted model were defined as true integrated utility neurons.

Before constructing the unrestricted and restricted models, we determined the preferred side of each neuron by fitting their activity with the utility and decision model, including left utility, right utility and decision as regressors (Supplementary Table 2.2) and defining the preferred side as the one in which its utility regressor has the highest absolute t-score.

Then, using the spike count Y of each neuron we fit an unrestricted GLM including q-values and uncertainty bonuses. We performed the model fitting and obtained a log-likelihood L_u using MATLAB's function $fitglm$:

$$\log(E(Y|x)) = b_0 + b_1 Q_{preferred} + b_2 B_{preferred} + b_3 Decision \tag{2.18}$$

To each neuron in this sub-population we also fit a restricted GLM including q-values but not uncertainty bonuses and obtained its log-likelihood L_r:

$$\log(E(Y|x)) = b_0 + b_1 Q_{preferred} + b_2 Decision \tag{2.19}$$

Finally, we performed likelihood ratio tests, with MATLAB's function $lratiotest$, between the unrestricted and restricted models, by computing the likelihood ratio test statistic $LR = 2(L_u - L_r)$, and comparing it to a chi-squared null distribution for

LR with one degree of freedom, stemming from one variable restriction. Neurons that rejected the null restricted model at a significance level of $\alpha = 0.05$ were defined as integrated utility neurons.

For the sub-population of integrated utility neurons, we used their fits from the unrestricted models to determine whether activity in these neurons correlated with q-values and uncertainty bonuses individually more than expected by chance. We averaged absolute t-scores for q-value and uncertainty bonus across integrated utility neurons to measure their collective degree of correlation regardless of excitation or inhibition. We then compared these values with average absolute t-scores obtained from bootstrapping 500 iterations of unrestricted model fits shuffling spike counts Y. We derived p-values from the number of times the true average absolute t-score surpassed the bootstrapped iterations.

Similarly, we performed a likelihood ratio test to test whether neurons encoded an integrated selected utility signal in the pre-decision period by fitting the following unrestricted model:

$$\log(E(Y|x)) = b_0 + b_1 Q_{selected} + b_2 B_{selected} \tag{2.20}$$

Subsequently, we compared the unrestricted model with the following null restricted model:

$$\log(E(Y|x)) = b_0 + b_1 Q_{selected} \tag{2.21}$$

We then followed the same likelihood test protocol described above to determine whether neurons would be classified as integrated utility neurons or not.

Dimensionality reduction and decoding with dPCA

To decompose the contribution of variables of interest and decisions to the neural population data and decode these variables interest from patterns of neural activity, we employed demixed principal component analyses (dPCA) (Kobak et al., 2016).

For each variable of interest, and each brain area, we created a pseudopopulation aggregating trials from all patients in order to generate a full data matrix X, with dimensions $(N, SQTK)$, where N is the total number of neurons recorded in that brain area, S is the number of stimuli quantiles used to partition trials (3: low, medium, and high), Q is the number of possible decisions (2: left and right), T is the number of time bins, and K is the number of trials used to construct the

pseudopopulation as described further. Firstly, we binned spike counts into 500ms bins, with a 16ms time window step. We repeated the binning procedure in two different time periods: the trial onset period (0s,2s), aligned to trial onset; and the pre-decision period (-2s,1s), aligned to button press.

Constructing pseudopopulations

To create neural pseudopopulations for dPCA, we pooled trials from all sessions and treated them as if they had been recorded simultaneously. To allow for trials from different sessions to be grouped together, despite having continuous variables of interest, we pooled groups of trials into 3 quantiles with the same number of trials, dividing the full range of each variable for each session into low, medium and high levels. After obtaining these quantiles, we assigned every trial in each session to one out of $3 \cdot 2 = 6$ categories, to account for all possible combination of quantile levels and decisions, and randomly sampled an equal number k of trials from each category, for each session, such that $\sum_{sessions} k = K$. We chose $k = 15$, for it to be small enough to allow sampling an equal number of trials from each of the 6 categories for every session, while including as many training examples as possible.

To mitigate biases introduced during the random trial sampling procedure, we repeated these steps 10 times, yielding 10 pseudopopulations, on which the dimensionality reduction and decoding procedures were repeated independently.

dPCA dimensionality reduction

For dPCA dimensionality reduction, the full data matrix X is centered over each neuron and decomposed as a factorial ANOVA, where t, s, and d are labels to indicate the time, stimulus, and decision marginalizations, respectively:

$$X = X_t + X_{ts} + X_{td} + X_{tsd} + X_{noise} = \sum_{\phi} X_\phi + X_{noise} \tag{2.22}$$

The goal of dPCA is then to minimize the regularized loss function, where F indicates the Frobenius norm and μ is the ridge regression regularization parameter, determined optimally through cross-validation:

$$L = \sum_{\phi} (||X_\phi - F_\phi D_\phi X||_F^2 + \mu ||F_\phi D_\phi||_F^2) \tag{2.23}$$

F_ϕ and D_ϕ are the non-orthogonal encoder and decoder matrices, respectively, arbitrarily chosen to have 3 components for each marginalization. Therefore, dPCA

aims to reduce the distance between each marginalized data set and their reconstructed version obtained by projecting the full data matrix onto a low-dimensional space with the decoders D and reconstructing it with the encoders F.

dPCA decoding

We used the same dPCA framework to perform population decoding of the variables of interest. The dPCA linear decoding pipeline has been previously described in detail (Kobak et al., 2016), but we will briefly discuss it here.

Firstly, the pseudopopulation data matrix X of dimensions $(N, SQTK)$ is divided into train and test datasets by leaving out one random trial for each of the SQ possible combinations of stimulus levels and chosen side, for all neurons and time points, to form X_{test} of dimensions (N, SQT) and X_{train} with the remaining data points. We perform this random trial sampling procedure 100 times for each of the 10 random pseudopopulations, resulting in a total of 1000 random resamples.

We performed the aforementioned dPCA steps with the train data matrix X_{train} to obtain a decoder matrix $D_{\phi,i}$, with $i = (1, 2, 3)$ representing each of the three demixed principal components for each marginalization ϕ.

To perform stimulus decoding, we iterate over the three components $i = (1, 2, 3)$, to obtain the mean projections over all train trials, for each stimulus class $s = (1, ..., S)$ pertaining to the current marginalization, and the vectors of decoded projections for test trials, for each unique test trial $k = 1, ..., SQ$, representing all the possible stimulus-decision combinations:

$$P_{\phi,s}^{train} = \begin{bmatrix} < D_{\phi,1} X_{train} >_s \\ < D_{\phi,2} X_{train} >_s \\ < D_{\phi,3} X_{train} >_s \end{bmatrix}, P_{\phi}^{test,k} = \begin{bmatrix} D_{\phi,1} X_{test,k} \\ D_{\phi,2} X_{test,k} \\ D_{\phi,3} X_{test,k} \end{bmatrix} \tag{2.24}$$

We then defined the decoded class C_k to be the one which minimizes the three-dimensional Euclidean distance between the test projection and the mean train projections:

$$C_k = \arg\min_s ||P_{\phi,s}^{train} - P_{\phi}^{test,k}|| \tag{2.25}$$

We obtained classification accuracy values for each trial resample by counting how many test trials were correctly labeled, and averaged classification accuracy

values over the 100 random test trial resamples, as well as the 10 pseudopopulation resamples.

Equivalently, to perform decision decoding, we follow the same steps, except that we obtain mean projections over all train trials for each decision class $q = (1, ..., Q)$ to compare with test trial projections.

Significance scores for each time bin were determined by obtaining the distribution of null scores from the random test trial reshuffles, and computing the quantile placement of the true decoding accuracy, assuming an approximate normal distribution for reshuffled decoding accuracies. We subsequently Bonferroni corrected significance scores for multiple comparisons across time bins.

dPCA component projection distance

To summarize how dPCA representations of utility and decision differ for low/high utility trials, as well as left/right decisions (Fig. 2.4E), for every time bin, we projected data X_{subset} from each trial subset (low utility trials, high utility trials, left decision trials, and right decision trials) onto the demixed principal components, expressed by the decoder matrix D, obtaining DX_{subset}. Note that each row of D represents one demixed principal component for the dataset. We then computed Euclidean distances between projections $D_{decision} = ||DX_{left} - DX_{right}||^2$, and $D_{utility} = ||DX_{high} - DX_{low}||^2$. We subsequently normalized projection distances into the [0,1] range.

Acknowledgements: We would like to thank the members of the O'Doherty and Rutishauser labs for discussions and feedback. We thank all subjects and their families for their participation, as well as nurses and medical staff for their work. This work was supported by National Institutes of Health Grants R01DA040011 and R01MH111425 (to J.P.O.), R01MH110831 (to U.R.), and P50MH094258 (to J.P.O. and U.R.).

Author Contributions: T.G.A., J.C., U.R., and J.P.O. designed the study. T.G.A. performed the experiments. T.G.A. and J.C. analyzed the data. T.G.A., J.C., A.N.M. U.R., and J.P.O. wrote the paper. A.N.M. performed surgery and supervised clinical work.

2.6 Supplementary Material

Model comparison

We compared how well two different computational models explained the observed behavior. Both models allowed for uncertainty and novelty to contribute to decision making beyond what could be implemented with a simpler reinforcement learning framework and also allowed incorporating patients' individual preferences. The compared models had two distinct mechanisms for how novelty, uncertainty, and q-value interacted to create stimulus utility (see Materials and Methods for detailed model descriptions). The first model we tested is a familiarity modulated upper confidence bound (fmUCB) model, shown to explain behavior from a neurotypical population in this task well (Cockburn et al., 2021) (Fig. 2.1G). In this model, stimulus utility is equal to a linear combination of the stimulus q-value and an uncertainty bonus, defined as a weighted product between uncertainty and novelty, allowing for the uncertainty and novelty factors to interact. The second model we tested was a 'linear novelty model', in which stimulus utility is equal to a linear combination of q-value, an uncertainty bonus derived from stimulus uncertainty alone, and a novelty bonus derived from stimulus novelty alone, without direct interactions between novelty and uncertainty.

We performed model fitting and comparisons for the two models across the patient population using hierarchical Bayesian inference (Piray et al., 2019) (see Fig. 2.7 for values of fit parameters). The fmUCB model explained the observed behavior across the patient population significantly better, with an estimated model frequency of 78.3% and exceedance probability of 99.6%. This is consistent with the results of a study from a larger cohort of healthy participants performing the same task (Cockburn et al., 2021), suggesting a non-linear interaction between uncertainty and novelty can drive decisions.

To test whether patient behavior reflected the instruction to reset the q-values and uncertainty bonuses at the beginning of every block as instructed, we also compared the fmUCB model with a 'no reset' model. This model was similar to the fmUCB model except that it did not reset q-value and uncertainty estimates from one block to the next. The fmUCB model had an exceedance probability of 99.9% compared to the no-reset model, indicating that patients reset contingencies between blocks as instructed. Lastly, we also compared the fmUCB model with a simpler reinforcement learning model in how well they could recover decision behavior. This analysis showed that a simple RL model did not adequately capture the observed behavioral

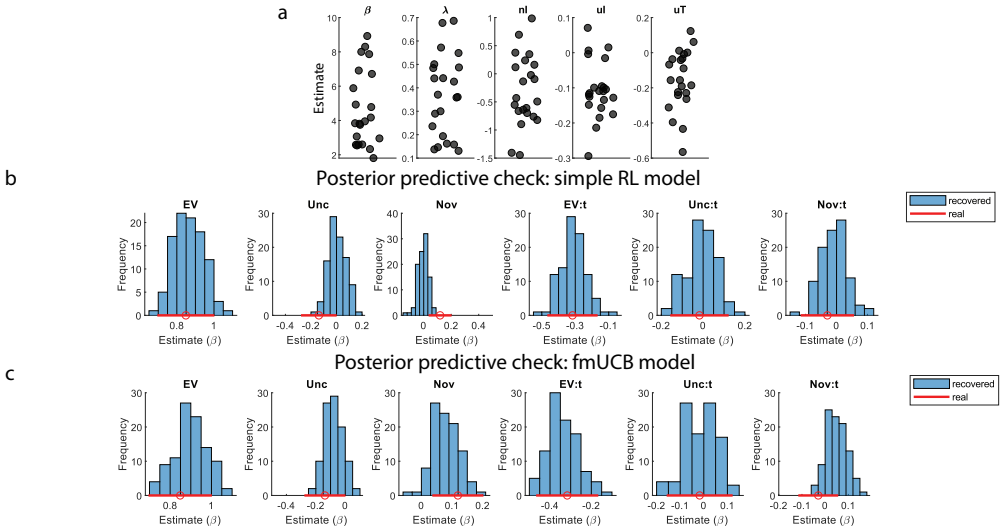

Figure 2.7: Model fits and posterior predictive check for selected exploration model with familiarity gating mechanism. (a) Individual fmUCB model parameter fits. Each dot represents a parameter fit for each patient (Left to right: softmax inverse temperature β; learning rate λ; novelty intercept; uncertainty intercept; uncertainty terminal). (b) Posterior predictive check for a simple reinforcement learning model which only included a softmax beta and a learning rate as free parameters. We fit this model to patient behavior and re-exposed an artificial agent with the obtained model parameters to the same set of trials which patients experienced 50 times, to generate decisions according to the estimated decision probabilities. We then fit a logistic regression for the effect of each variable (left to right: expected values, uncertainty, novelty, and their respective interactions with trial number) on decision in the artificial agents (blue histogram) and compared it to the actual estimate given true decisions concatenated across patients (red line; dot indicates mean and bars indicate 95% confidence interval). (c) Same, for the fmUCB model, which was selected for all subsequent analyses.

effects beyond seeking expected value.

These behavioral modeling results show that a fmUCB mechanism is appropriate to model the behavior of our subjects. For all subsequent neural data analysis we therefore used the variables derived from the fmUCB model as regressors.

Additional behavioral analysis

For the chosen fmUCB model, we summarized its parameter fits to gain more insight into the behavior of the patients as a group (Fig. 2.7A). The softmax inverse temperature was 4.6 ± 0.47 and the learning rate was 0.36 ± 0.03. The novelty intercept parameter represents a novelty initiation (nI) bias and was negative (-0.30 ± 0.13, $p < 0.05$, t-test), indicating a slight preference for familiar stimuli in the early

trials of a block. The uncertainty intercept (uI) parameter represents how much value was assigned to uncertainty in the beginning of a block. This parameter was negative (-0.10 ± 0.01, $p < 10^{-5}$, t-test), indicating a slight uncertainty avoidance early in the block. The uncertainty terminal value parameter, which represents the value assigned to uncertainty at the end of a block, was also negative (-0.15 ± 0.03, $p < 10^{-4}$, t-test), indicating that uncertainty was still valued negatively on average by the end of the blocks.

To ensure that the fmUCB model correctly reproduces the effects of expected value, uncertainty and novelty observed in our subjects, we performed a posterior predictive check analysis (Fig. 2.7B-C). We exposed the fmUCB model to the same sequence of trials each patient experienced, with each patient's model fits, and generated a decision for each trial used the decision probabilities inferred from the model. We then used these choices to fit a logistic regression which maps the effects of the various variables on the choices (equivalent to Fig. 2.1F). To account for variability in probabilistic decisions, we repeated this procedure 100 times and generated a distribution of regression coefficient estimates, which we compared to the actual effects observed in the subjects. For comparison, we performed the same procedure with an equivalent model (a simple reinforcement learning model), except that it did not model any effects of uncertainty and novelty, including only the softmax beta and learning rate parameters. To quantify the difference in how these models recovered effects of expected value, uncertainty and novelty on behavior, we computed the 95% confidence intervals for the logistic regression estimates of value feature effects on actual patient decisions (as displayed in Fig. 2.1, see Materials and Methods). Then, we obtained the proportion of overlap between these confidence intervals and the recovered effect estimates obtained from 100 iterations of simulated sessions in the posterior predictive check analysis. Using the simple RL model, we observed an overlap of $[95\%, 50\%, 14\%]$ for the estimates of expected, uncertainty and novelty, respectively. With the fmUCB model, we obtained overlaps of $[89\%, 88\%, 85\%]$. Therefore, the fmUCB model captured the behavioral effects of uncertainty and novelty better, while a simpler reinforcement learning model was not appropriate to capture the observed behavioral effects beyond seeking expected value.

Encoding rejected values

Models of decision making often also depend on maintaining information about the value of options not chosen to evaluate the outcome received. We therefore also examined whether rejected q-values, uncertainty bonuses, and novelties were

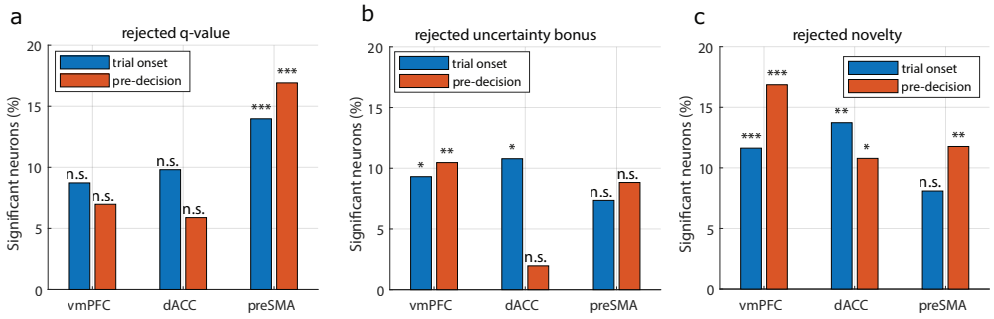

Figure 2.8: Single neuron encoding for the q-value, uncertainty bonus and novelty of the rejected stimulus in each trial. (a) Proportion of neurons sensitive to rejected q-value in vmPFC, dACC, and preSMA, in the trial onset (blue) and pre-decision (orange) periods. Stars indicate neuron count significance in a binomial test (* = p<0.05; ** = p<0.01; *** = p<0.001, Bonferroni corrected).

represented (Fig. 2.8A-C). Unlike selected q-values, preSMA did represent rejected q-values (trial onset 14.0%, $p < 10^{-4}/6$:, pre-decision: 16.9%, $p < 10^{-6}/6$, binomial test, Bonferroni corrected). Novelty of the rejected option was encoded in vmPFC (trial onset 11.6%, $p < 0.001/6$:, pre-decision: 16.8%, $p < 10^{-7}/6$, binomial test, Bonferroni corrected).

Exploratory signaling in vmPFC uncertainty neurons

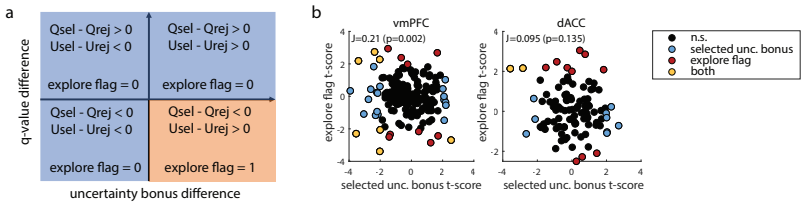

Figure 2.9: Comparing encoding of selected uncertainty bonus and exploration. (a) Chart indicating how trials were defined as explore or non-explore trials. Trials in which the selected option had lower q-value and higher uncertainty bonus were defined as explore trials (orange) and all other trials were defined as non-explore trials (blue). (b) Scatter plot of selected uncertainty t-scores versus explore flag t-scores from the Poisson GLM analysis. We display neurons sensitive to selected uncertainty bonus (blue), to the explore flag (red), and to both regressors (yellow). We also indicate the Jaccard overlap index and a p-value for the Jaccard test, indicating a significant overlap between uncertainty and exploration coding in individual vmPFC neurons. Left: vmPFC neurons; Right: dACC neurons.

One potential reason for representing the uncertainty of the selected option is to enable exploratory decision making, which would entail deliberately choosing an

item with lower q-value. We therefore divided trials into putative explore and non-explore categories. Trials in which the patient chose the option which had the lower q-value but the higher uncertainty bonus were classified as putative explore trials, while all others were classified as non-explore trials (Fig. 2.9A).

Then, in the sub-populations of vmPFC and dACC neurons in the pre-decision period that were sensitive to selected uncertainty, we performed a Poisson GLM analysis using the explore trial flag as a regressor, correcting for selected uncertainty as a regressor of no interest. We subsequently tested whether neurons whose activity was significantly modulated by the explore flag significantly overlapped with the sub-populations of vmPFC and dACC neurons that encoded selected uncertainty (Fig. 2.9,B). We found a significant overlap in vmPFC ($p < 0.01$, Jaccard index test), but not in dACC ($p = 0.13$, Jaccard index test). Therefore, a significant proportion of vmPFC selected uncertainty neurons signal whether a trial is exploratory or not prior to button press.

2.7 Supplementary Tables

Task version	Patients who performed it
Longer (300 trials)	P60,P61,P62,P63,P64,P65,P67,P69,P70,P71
Shorter (206 trials)	P41,P43,P48,P49,P51,P54,P55,P56

Table 2.1: Patients who performed the longer (300 trials) or shorter (206 trials) version of the task. For all behavioral and neural analyses, datasets from both task versions were pooled.

Name	Model	Periods
Full positional model	$\log(E(Y\|x)) = b_0 + b_1 Q_L + b_2 Q_R + b_3 B_L + b_4 B_R + b_5 N_L + b_6 N_R$	trial onset, pre-decision
Full selection-based model	$\log(E(Y\|x)) = b_0 + b_1 Q_{sel} + b_2 Q_{rej} + b_3 B_{sel} + b_4 B_{rej} + b_5 N_{sel} + b_6 N_{rej}$	trial onset, pre-decision
Selection-based utility model	$\log(E(Y\|x)) = b_0 + b_1 U_{sel} + b_2 U_{rej}$	trial onset, pre-decision
Explore flag model	$\log(E(Y\|x)) = b_0 + b_1 Explore + b_2 U_{sel}$	pre-decision
Decision and utility model	$\log(E(Y\|x)) = b_0 + b_1 U_L + b_2 U_R + b_3 Decision$	trial onset, pre-decision
Outcome model	$\log(E(Y\|x)) = b_0 + b_1 O + b_2 Q_{sel} + b_3 \|RPE\|$	outcome

Table 2.2: Models for Poisson GLM single neuron encoding analysis.

Patient ID	Left dACC	Right dACC	Left preSMA	Right preSMA	Left vmPFC	Right vmPFC
P71CS	-1.00,27.97,27.02	7.68,28.87,23.68	-1.89,13.51,38.03	5.09,11.30,39.61	-3.04,32.45,-0.64	5.06,34.29,6.30
P70CS	-0.84,23.06,26.04	4.25,22.06,22.63	-9.45,11.24,46.67	12.40,11.55,46.60	-0.20,26.30,-6.79	10.50,29.65,-3.63
P69CS	-6.13,27.44,22.69	-1.66,27.37,22.25	-2.02,23.89,39.02	2.92,32.34,40.50	-3.73,38.30,-11.69	2.26,29.24,-9.80
P67CS	-4.82,27.63,-11.69	-0.27,28.18,-9.13	-2.63,10.61,37.48	7.38,11.36,36.23	-4.70,27.32,-14.06	-1.21,28.12,-7.68
P65CS	-4.66, 24.50, 24.48	5.66, 30.60, 25.28	-0.82, 10.87, 52.78	1.53, 18.10, 48.52	-3.69, 45.06, -12.11	6.46, 30.94, -11.70
P64CS	-7.45, 35.50, 22.73	5.54, 32.54, 30.78	-6.19, 21.34, 47.67	7.46, 22.79, 44.56	-8.18, 30.42, -13.39	6.21, 31.14, -12.75
P63CS	-5.57, 27.44, 26.20	11.66, 34.65, 33.43	-8.53, 10.02, 41.57	9.46, 28.48, 48.52	-3.52, 45.04, -14.44	9.54, 37.31, -13.49
P62CS	NaN	9.02, 15.72, 29.33	NaN	5.69, 16.28, 49.18	NaN	11.02, 34.13, -10.05
P61CS	0.00, 19.15, 22.74	6.82, 20.21, 27.63	-4.55, 14.50, 38.70	14.92, 12.05, 36.59	-3.40,41.84,-15.66	6.99,31.03,-12.39
P60CS	-8.66, 30.23, 27.13	5.62, 31.09, 20.05	-8.70, 19.60, 44.90	5.55, 26.23, 42.00	-1.49,31.63,-16.97	2.48,33.28,-20.24
P56CS	-7.01, 23.11, 30.17	16.27, 24.79, 25.90	-6.92, 6.43, 46.70	29.08, 13.01, 45.88	-6.96,34.62,-13.05	27.02, 31.69, -5.48
P55CS	-4.14, 28.97, 24.60	3.20, 28.05, 21.05	-3.88, 11.08, 46.18	7.58, 13.24, 42.29	-8.64, 36.74, -6.98	6.32, 35.04, -11.70
P54CS	-3.85, 29.02, 23.70	7.93, 33.96, 26.74	-5.98, 20.87, 49.08	10.72, 26.92, 48.06	-3.53, 36.85, -16.95	3.94, 31.75, -14.51
P53CS	0.95, 32.02, 27.76	9.47, 36.81, 26.07	-4.30, 12.32, 51.06	10.21, 18.90, 49.06	-0.97, 35.20, -17.12	6.83, 31.74, -14.96
P51CS	-5.48, 31.63, 27.98	6.65, 30.73, 25.95	-8.81, 8.15, 44.07	11.91, 9.98, 47.01	-1.03, 36.31, -10.91	4.19,30.23,-15.23
P49CS	-8.72, 34.10, 25.92	6.21, 31.04, 25.99	-5.54, 25.16, 31.55	6.86, 26.19, 38.56	-7.02, 36.94, -13.11	10.87, 35.06, -5.09
P48CS	-3.34, 28.54, 24.20	5.87, 40.15, 17.96	-7.58, 24.03, 37.93	11.91, 23.85, 40.98	-8.06, 36.04, -6.05	13.66, 37.28, -17.09
P47CS	-0.67, 25.84, 20.14	1.43, 25.49, 17.20	-7.86, 12.20, 42.04	8.11, 11.45, 45.05	-1.98, 40.24, -15.36	1.02,38.60,-13.11
P43CS	-2.85, 23.88, 24.58	5.29, 27.07, 24.76	-2.83, 9.40, 43.93	5.33, 8.85, 46.26	-4.26, 38.19, -12.07	5.11,38.48,-16.47
P41CS	0.18, 26.56, 17.92	12.67, 27.14, 19.01	-0.04, 12.88, 58.69	6.99, 15.53, 57.30	-3.17, 37.28, -7.56	8.67, 38.04, -15.47

Table 2.3: MNI coordinates for microelectrode positioning in all patients in dACC, preSMA and vmPFC.

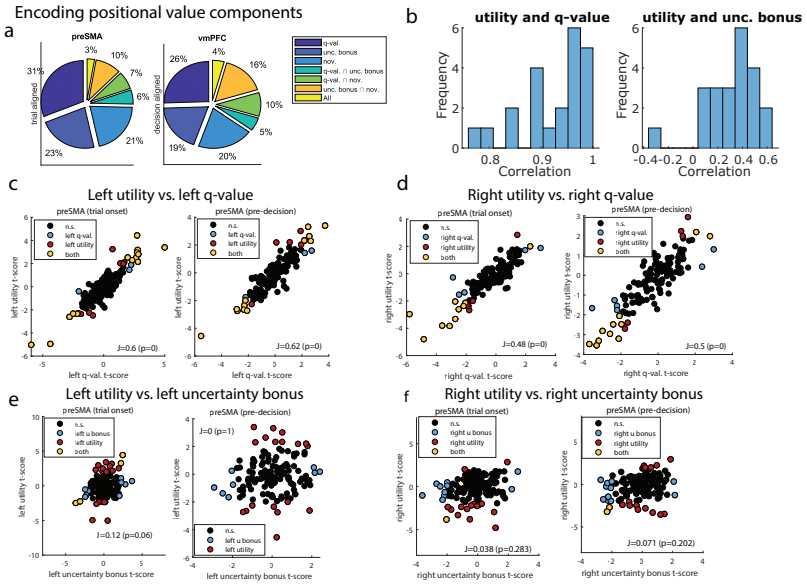

Figure 2.10: Summarizing positional encoding and comparing encoding of utility versus value components. (a) Pie chart including proportion of neurons sensitive to each positional component of value, and their respective overlaps, in preSMA (left) and vmPFC (right). (b) Histogram of correlation between utility and q-value (left), or utility and uncertainty bonus (right) trial vectors, across recording sessions. (c) Given the sizeable correlations between utility and its components, we mapped out the extent to which left utility preSMA neurons also correlated with left q-value preSMA neurons, in the trial onset period (left), and the pre-decision period (right). We plot q-value t-scores from the positional components GLM, as well as the utility t-scores from the utility and decision GLM (black: non-sensitive neurons; blue: left q-value neurons; red: left utility neurons; yellow: neurons concurrently sensitive for both). We tested whether this overlap is significant and report the Jaccard index (J), as well as p-values from the Jaccard test of overlap. (d) Same, for right utility vs. right q-value. (e) Same, for left utility vs left uncertainty bonus. (f) Same, for right utility vs. right uncertainty bonus.

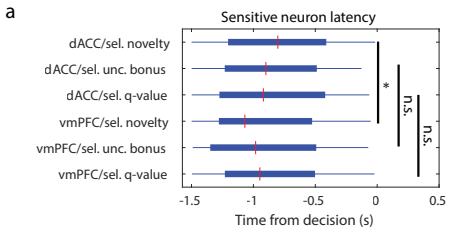

Figure 2.11: Timing for neurons which encode selected components of value. (a) Box plots of latency time across trials for all selected q-value, selected uncertainty bonus, or selected novelty neurons in vmPFC/dACC. The red mark indicates the median, and the box extends between the 25th and 75th percentiles of latency times. Bar whiskers extend to the most extreme data points not labeled as outliers, defined as values that are more than 1.5 times the interquartile length away from the edges of the box. Stars indicate significance in a two-sided rank-sum test between latencies for each regressor (* = $p<0.05$).

Chapter 3

SINGLE NEURON CORRELATES OF MODEL-BASED PAVLOVIAN CONDITIONING IN THE HUMAN BRAIN

3.1 Abstract

Despite behavioral evidence in favor of cognitive map acquisition during Pavlovian learning, most computational accounts of classical conditioning have relied on model-free mechanisms to explain neural and behavioral data. In this study, we leveraged human single unit recordings in ventromedial prefrontal cortex, amygdala and other brain areas to investigate stimulus-stimulus associations and identity based coding, which are components of a model-based learning framework. We found evidence of stimulus-stimulus associations in vmPFC, while both vmPFC and amygdala performed predictive value coding. Additionally, we found that the temporal correlations between vmPFC and amygdala spikes was modulated by the expected value of conditioned stimuli. These results shed light on the formation of cognitive maps during Pavlovian conditioning in the human brain.

3.2 Introduction

One of the most fundamental aspects of learning is the ability to create predictive associations between stimuli and outcomes. In Pavlovian conditioning, also known as classical conditioning, the associations that an organism creates can be leveraged to produce adaptive anticipatory behaviors, such as saliva production following presentation of a cue which is predictive of receiving a food reward, or avoidance following a cue paired with an aversive or threatening stimulus (Pavlov, 1927; R. A. Rescorla, 1988; O'Doherty, Cockburn, and Pauli, 2017).

Pavlovian conditioning differs from other forms of associative learning, such as instrumental conditioning, in that it is strictly defined by associating stimuli and outcomes passively, independently from the animal's behavior, though it may require active cognitive processing of the relationships between stimuli to occur (R. Rescorla, 1988). Under this definition, Pavlovian conditioning was originally described in dogs (Pavlov, 1927), but has since been observed in numerous other animals, such as *Schistocerca* (grasshopper) (Dukas and Bernays, 2000), *Drosophila* (fruit fly) (Tully and Quinn, 1985), *Aplysia* (sea slug), (Walters, Carew, and Kandel, 1981), pigeons (Brown and Jenkins, 1968), and rats (R. A. Rescorla, 1988). In

humans, the role of Pavlovian conditioning has been described in processes such as phobias (Davey, 1992) and addiction (Poulos, Hinson, and Siegel, 1981), while learning theories derived from Pavlovian conditioning, such as the Rescorla-Wagner model (R. Rescorla and Wagner, 1972), have been successfully adapted to describe a variety of associative learning processes (Siegel and Allan, 1996).

Additionally, Pavlovian conditioning behavior exhibits a number of core properties, all of which have been documented in human studies: latent inhibition, which consists in hampered conditioning when a stimulus-outcome pairing is attempted after the stimulus has been previously presented without the outcome (Siddle, Remington, and Churchill, 1985; Lubow and Gewirtz, 1995); blocking, which inhibits the association between a new stimulus and an outcome when the new stimulus is presented alongside another cue which is already fully predictive for the outcome (Arcediano, Matute, and R. R. Miller, 1997); sensory pre-conditioning, in which stimulus A can start eliciting a conditioned response if it had been previously paired with an initially neutral stimulus B, when stimulus B itself is subsequently associated with the outcome (Brogden, 1947; White and Davey, 1989); and higher order conditioning, which allows for stimuli to elicit a conditioned response when they are presented in a sequence with other stimuli which are themselves predictive of the outcome (Seymour et al., 2004; Pauli, Larsen, et al., 2015; Pauli, Gentile, et al., 2019). Ultimately, a unifying framework for these properties is that Pavlovian associative learning hinges on how informative a stimulus is about the probability of a subsequent outcome.

Computational theories of Pavlovian conditioning have been derived from models of instrumental conditioning, most prominently through the class of model-free (MF) reinforcement learning models (Daw, Niv, and Dayan, 2005; O'Doherty, Cockburn, and Pauli, 2017). In this framework, stimulus values are learned based on stimulus-outcome or action-outcome associations alone, without depending on a cognitive map for the transition structure between stimuli. This class of models can be thought of as developments from Thorndike's law of effect, stating that rewarded actions are more likely to be repeated, while punished actions are more likely to be avoided (Thorndike, 1898). However, a learning framework which solely employs stimulus-outcome associations is insufficient to explain Pavlovian phenomena such as sensory pre-conditioning, which requires stimulus-stimulus associations to occur, independent of outcome presentation, as well as sensitivity to revaluation or devaluation, which may occur regardless of new outcome pairings

(Dayan and Berridge, 2014; Pool et al., 2019). Such phenomena can be modeled more accurately by a model-based (MB) account of learning, in which a internal cognitive map (Tolman, 1948) is developed to provide the agent with an internalized transition structure describing the probabilities of moving between states, requiring that the identities of stimuli and outcomes are tracked. In this framework, an agent can seek out stimuli which lead to newly valued outcome states even before pairing these stimuli with rewards, as long as the transition probabilities between states are known. In a purely model-free learning framework, on the other hand, an agent would need to be re-exposed to newly valued outcomes, and only then it would be possible to update the values of the cues associated with these outcomes.

Even though behavioral evidence suggests the model-based framework can provide an accurate account of Pavlovian conditioning, most previous work has focused on model-free mechanisms such as the Rescorla-Wagner model (R. Rescorla and Wagner, 1972), and temporal difference (TD) learning (Sutton, 1988), already yielding valuable insight about the neural implementation of learning processes. For instance, in TD models, the learning signal which updates stimulus values following outcome is a reward prediction error signal (RPE), which has been found to correlate with the activity of dopaminergic neurons in the ventral tegmental area (VTA) and the substantia nigra of rats (Schultz, Dayan, and Montague, 1997), and with BOLD signal in the human ventral striatum (O'Doherty, Dayan, et al., 2003), as obtained with fMRI.

Still, more recent studies have been starting to map how model-based Pavlovian conditioning occurs in the brain. In rats, a Pavlovian paradigm revealed that intact activity in both ventral striatum and orbitofrontal cortex (OFC) was necessary for model-based learning to take place (McDannald et al., 2011). Another rat study found that previously unpleasant Pavlovian cues associated with a salty stimulus could instantly become appetitive when the animal encountered them in a state of sodium depletion (M. J. Robinson and Berridge, 2013). This behavioral change, which is consistent with model-based Pavlovian conditioning, was accompanied by an increase in Fos activation in a mesocorticolimbic circuit including VTA, nucleus accumbens and OFC.

In humans, an fMRI study found correlations between amygdala activity and components of a model-based Pavlovian inference model (Prévost, McNamee, et al., 2013). Specifically, model-based estimates of a cue's expected value (EV) correlated with activity in basolateral amygdala (BLA) during an appetitive session,

and correlated with activity in the centromedial complex of the amygdala during an aversive session. Another human study investigated the neural representation of stimulus-stimulus associations which could be a substrate for model-based Pavlovian conditioning (Pauli, Gentile, et al., 2019). This study used a sequence of two cues (distal and proximal) which had a probabilistic transition structure, followed by an appetitive or neutral outcome. The authors found that the decoding accuracy for stimulus identity in caudate nucleus correlated with the explicit knowledge that participants had about stimulus-stimulus associations. Crucially, a classifier trained using OFC activity to decode the identity of proximal cues during proximal cue presentation performed better than chance when tested during distal cue presentation, indicating that OFC already encoded predictive information about the identity of the proximal cue at the time of the distal cue, which suggests a neural substrate for stimulus-stimulus associations. Other studies in humans also found a link between OFC activity and outcome identity representation during reward learning (Klein-Flügge et al., 2013; Howard et al., 2015).

In this study, we leveraged single neuron recordings in patients undergoing treatment for refractory epilepsy to investigate a number of open questions on how model-based Pavlovian conditioning occurs in the human brain. Specifically, is there evidence for encoding of stimulus-stimulus associations and stimulus identities in vmPFC neurons, which are fundamental in the construction of cognitive maps? Can we map how amygdala neurons act in tandem with prefrontal neurons in predictive value coding, which is a key feature of Pavlovian conditioning? Additionally, how is outcome feedback encoded, alongside with the surprise signals which are required to update cognitive maps during learning?

3.3 Results

Behavioral evidence of Pavlovian conditioning

We recorded 165 AMY, 119 HIP, 86 vmPFC, 137 preSMA, and 103 dACC single neurons (610 total) in 13 sessions from 12 patients chronically implanted with hybrid macro/micro electrodes for epilepsy monitoring. Patients performed a sequential Pavlovian conditioning task (Fig. 3.1A,C) with two conditioned stimuli in the form of fractal images: distal (CSd), followed by proximal (CSp). Conditioned stimuli were then followed by an outcome, which could be rewarding or neutral (Pauli, Larsen, et al., 2015; Pauli, Gentile, et al., 2019). Outcomes were delivered in the form of videos, either of a hand depositing a piece of candy in a bag, or of an empty hand approaching a bag. Patients were told that every display of the rewarding

video contributed partially to the amount of real candy they would be given after the end of the session. Patients were asked to pay attention to CS identities as they would be predictive of rewards and were told to perform a button press during CSp presentation as an attention check, but were informed this did not influence the outcome of the trial in any way. In each block, a CSd/CSp pair would be more likely associated with the reward, and were therefore defined as CSd+/CSp+ (see Materials and Methods for task details), according to a common/rare transition structure. Conversely, the other CSd/CSp pair was more likely to precede the neutral outcome, and are referred to as CSdn/CSpn.

We fit a normative model-based transition matrix model (see Materials and Methods for model details) to infer expected values (EV), state prediction errors (SPE) and transition probabilities on a trial by trial basis, for each session (Fig. 3.1E). With these transition probabilities, we could also estimate which CSp was most likely to follow a CSd in each trial, which we refer to as *CSp presumed identity*. To infer whether Pavlovian conditioning occurred across patients, we correlated the obtained model covariates with behavioral metrics such as stimulus ratings and pupil dilation.

Subjective preference ratings for all fractal images were obtained in the beginning of the task. Additionally, between blocks, we asked patients to re-rate the fractals that were included in the previous block to obtain measures of changes in subjective preference as a function of patients' experience in the task (Fig. 3.1B). When grouping CSd and CSp together, we observed that stimuli used as CS+ were rated significantly higher than stimuli used as CSn ($p = 0.02$, one sided t-test). We also tested whether distal and proximal stimuli had their ratings change by a different amount by contrasting absolute rating changes for (CSd+,CSdn) versus (CSp+,CSpn), and found no significant difference for distal vs. proximal stimuli ($p=0.16$, two-tailed t-test).

Pupil diameter was analyzed in two distinct time windows: during CSd presentation and CSp presentation (see Materials and Methods for details on pupil analysis). We obtained the average pupil diameter change within these periods, relative to a baseline, and tested whether they correlated with model covariates with a linear mixed effects model, with session number as a random effect. Specifically, the model for pupil diameter during CSd presentation included the EV of the distal CS, while the model for pupil diameter during CSp presentation included CSp EV, SPE, and an interaction term between CSp EV and SPE. We found no effect for CSd EV in the first model ($p = 0.30$). In the second model (Fig. 3.1D), we did not

find an effect for CSp EV (p = 0.07) or SPE (p = 0.06), but we did find an effect for the interaction term between CSp EV and SPE (p = 0.02), indicating that pupil diameter correlated with a combination of computational factors inferred from the model-based framework. A similar interaction result was previously observed in a Pavlovian conditioning paradigm performed in a neurotypical population (Pauli, Larsen, et al., 2015).

Overall, the aggregate behavioral evidence from changes in subjective stimulus ratings and pupil sensitivity to an interaction of EV and SPE suggests that conditioning took place across patients.

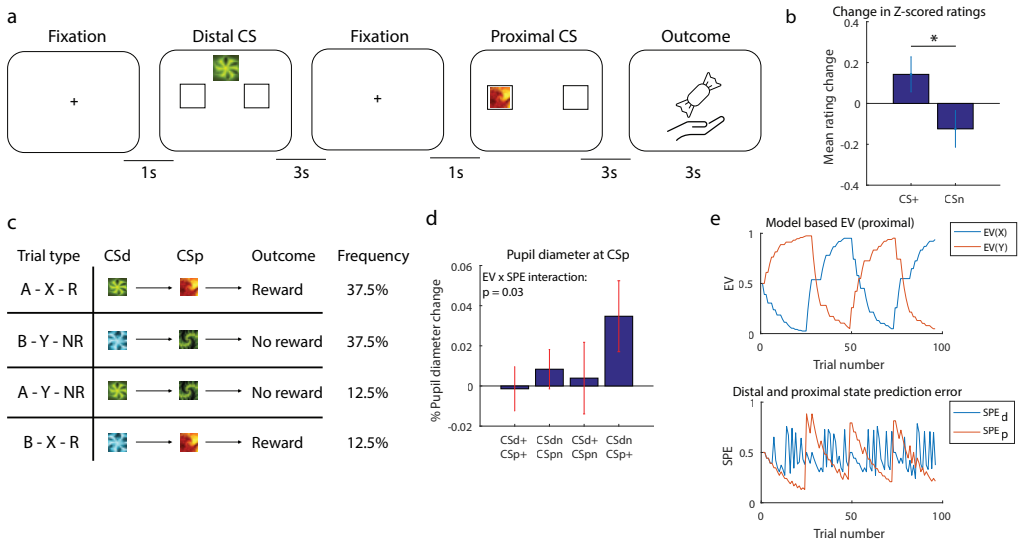

Figure 3.1: Pavlovian conditioning task and behavior. (a) Trial structure. After a fixation period, patients saw a sequence of two conditioned stimuli, distal and proximal, with a 1s fixation period in between them. Then, outcomes were presented: for positive outcomes, a video of a hand depositing a piece of candy in a bag; for neutral outcomes, an empty hand approaching a bag. (b) Changes in stimulus ratings. After each block, patients rated stimuli for their subjective preference. We compared how ratings changed for each fractal compared to its previous value, depending on whether they were a positive or neutral CS in that block. (c) Trial types. Stimuli transitioned from distal to proximal according to a common/rare probabilistic structure. The same 2 fractals were used as CSp throughout the entire task, while new fractals were picked as CSd in every block. (d) Pupil diameter change at CSp. We compared pupil diameter during CSp presentation with a baseline period, for each trial type. Error bars represent SEM. (e) Model-based regressor examples. Top: Model-based expected value for the two possible CSp. Bottom: Distal (blue) and proximal (orange) state prediction error in each trial.

Single neurons encode stimulus identity and model-based regressors

We next investigated whether firing rates in individual neurons correlated with task variables and the estimated model-based covariates, using a Poisson GLM analysis (see Materials and Methods for details, Fig. 3.2). We obtained spike counts for each neuron in the time windows that were relevant for each regressed variable (e.g. counting spikes during outcome presentation for regressing outcomes). After obtaining significance results for each neuron, we tested whether the number of significant neurons in each brain region was more than expected by chance, with a binomial test, Bonferroni corrected for the number of tested brain areas.

We regressed the CSp presumed identity at the time of CSd presentation (Fig. 3.2A), to test whether the most likely identity of the next presented stimulus was already encoded by neurons at distal time. We found that 11.6% of vmPFC neurons encoded CSp presumed identity ($p < 0.05/5$), indicating that vmPFC neural activity produces predictive identity representations in a stimulus-stimulus association context. This type of activity is necessary for model-based learning to take place in this sequential Pavlovian conditioning paradigm. An example neuron performing this type of encoding is shown in Fig. 3.2C. We also regressed the EV of the presumed CSp at distal time and the actual CSp identity at proximal time (Fig. 3.2B) but did not find a significant neuron count in any region.

At outcome time (Figs. 3.2E,F), we found a significant proportion of neurons correlated with outcome (reward vs. neutral) in hippocampus (11.8%, $p < 0.01/5$) and preSMA (12.4%, $p < 0.001/5$), as well as a significant proportion of neurons correlated with SPE in hippocampus (11.8%, $p < 0.01/5$), dACC (13.6%, $p < 0.001/5$), and preSMA (9.49%, $p < 0.05/5$). This indicates that neural activity in these areas not only distinguishes between rewarding and neutral outcomes but also tracks how surprising it was to arrive at each outcome state, which is a fundamental component of learning within the model-based framework. We also tested for correlations with SPE at CSp presentation time but did not find significant neuron counts in any brain region.

Populational decoding of identity and value

We next investigated if the joint activity patterns from neurons recorded simultaneously in each brain area were predictive of several variables of interest. For this, we performed a cross-validated populational decoding analysis with a linear SVM, obtaining significance levels with a bootstrapped null distribution (see Materials and

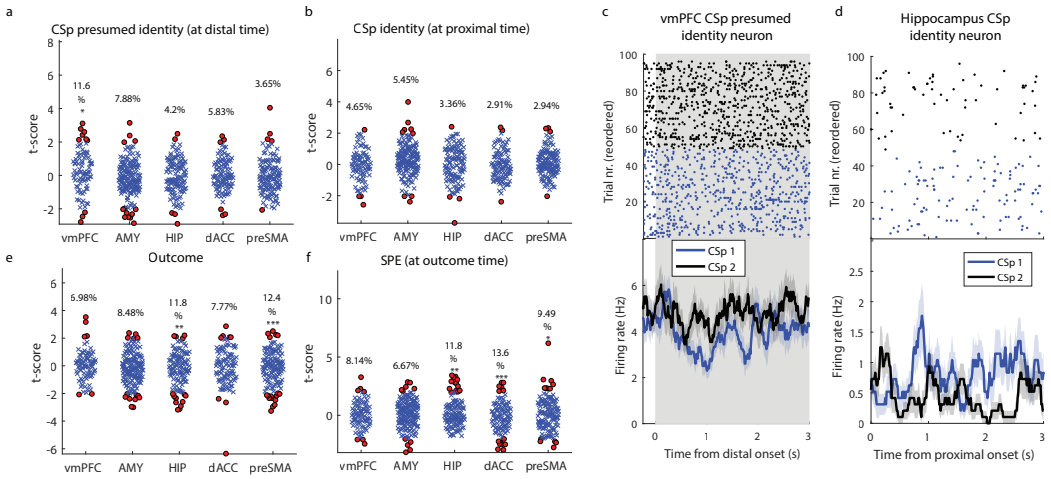

Figure 3.2: Single neuron encoding. (a) T-scores for every neuron in each brain area, for a GLM predicting spike counts during CSd presentation with the presumed identity of the CSp as a regressor. Red dots indicate significant neurons and stars indicate significance across the entire region, corrected across areas. (b) Same, for CSp identity during CSp presentation. (c) vmPFC neuron whose activity during distal presentation correlates with the presumed identity of the proximal stimulus (blue: first CSp; black: second CSp). Top: raster plot; Bottom: PSTH. (d) Same, for a CSp identity neuron in hippocampus, during proximal presentation. (e) Same as B, for outcome. (f) Same, for state prediction error at outcome time.

Methods for details).

We found that CSp presumed identity could already be significantly decoded at distal time in vmPFC ($p < 0.01/5$, permutation test), in consonance with the significant neuron count we found (Fig. 3.3A). This further establishes vmPFC neural activity as a substrate for predictive coding during stimulus-stimulus associations in model-based Pavlovian conditioning. During CSp presentation, however, CSp identity could only be significantly decoded in hippocampus (Fig. 3.3B), despite a lower-than-chance neuron count in encoding. This could reflect the contribution toward identity decoding accuracy from multiple neurons which were individually under significance threshold.

Similarly, we found significant decoding of outcomes in dACC ($p < 0.001/5$, permutation test), which had a sub-threshold neuron count in the encoding analysis. Moreover, in consonance with the encoding results, we found significant outcome decoding in hippocampus ($p < 0.05/5$, permutation test) and preSMA ($p < 0.01/5$, permutation test), highlighting a widespread representation of outcome valence. Importantly, we could decode model-based estimates of CSd EV at distal time

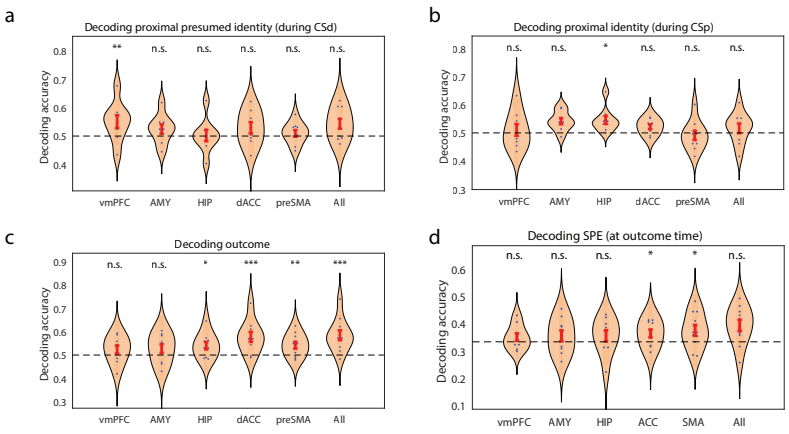

Figure 3.3: Decoding outcomes and stimulus identity. (a) Decoding accuracy for the identity of the presumed CSp during distal presentation. Each dot indicates accuracy in one session, stars indicate significance across sessions with a bootstrapped null distribution, corrected across areas. Bars and dashed lines indicate standard error and chance level, respectively. (b) Same, for proximal identity during CSp presentation (c) Same, for outcome during the outcome period (d) Same, for state prediction error during the outcome period.

in preSMA ($p < 0.05/5$, permutation test, Fig. 3.4A), as well as CSp EV at proximal time in vmPFC ($p < 0.01/5$, permutation test) and amygdala ($p < 0.01/5$, permutation test, Fig. 3.4A). Taken together, these results indicate that vmPFC performs not only predictive stimulus-stimulus coding, but also predictive value coding in consonance with amygdala, as dynamically estimated by a model-based learning framework.

Correlations between vmPFC and amygdala neurons are modulated by expected value

Given the robust populational decoding of model-based expected values for the proximal CS at proximal time that we found in vmPFC and amygdala, we tested the hypothesis that cross-correlations for vmPFC-amygdala neuron pairs were modulated by EV in this time period. For this, we split all trials in 3 EV tertiles and defined the first and third tertiles as low and high EV trials, respectively, discarding the middle tertile. Then, for the low and high EV trials, we separately computed spike-spike cross-correlations for all simultaneously recorded vmPFC-amygdala neuron pairs (see Materials and Methods for details). Finally, to summarize our results we computed cross-correlogram integrals, separately for the positive and negative time lag periods. A peak in the positive time lag region indicates that amygdala spikes pre-

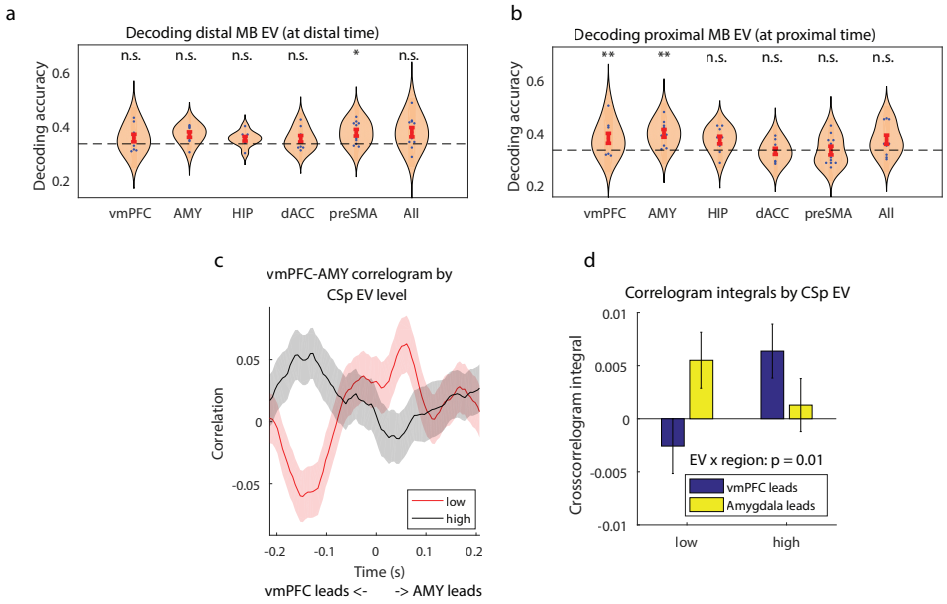

Figure 3.4: Decoding Pavlovian model-based value and neural cross-correlations. (a) Decoding accuracy for model-based EV during distal presentation. Each dot indicates accuracy in one session, stars indicate significance across sessions with a bootstrapped null distribution, corrected across areas. Bars and dashed lines indicate standard error and chance level, respectively. (b) Same, for model-based EV at proximal presentation. (c) Spike-spike cross-correlation between vmPFC and amygdala neurons recorded in the same sessions. Correlograms were computed separately by level of proximal model-based EV (red: low; black: high). (d) Correlogram integrals by level of proximal model-based EV (low or high). Integrals were computed as a summary metric, separately in the positive (amygdala leads, yellow) and negative (vmPFC leads, blue) time lag regions.

ceded vmPFC spikes more often, whereas a peak in the negative time lag region indicates that vmPFC spikes preceded amygdala spikes more often.

We found that the average cross-correlogram peaked at +60ms (amygdala leading) for low EV trials and peaked at -0.127ms (vmPFC leading) for high EV trials, indicating an inversion in the most likely leading region in spike correlations as a function of EV level (Fig. 3.4C). Additionally, an ANOVA on correlogram integrals, with leading region and EV level as factors revealed a significant interaction between these two factors ($p = 0.01$, Fig. 3.4D). Taken together, these results shed light on the relationship between amygdala and vmPFC in predictive value coding during Pavlovian conditioning. While both regions contained predictive information for model-based value, which region's spikes tended to precede the other's was determined by the expected valence of the outcome.

3.4 Discussion

In this study, we investigated the neural bases of model-based cognitive maps for Pavlovian conditioning, using single neuron recordings from the human brain during a sequential conditioning task. We found evidence for coding of stimulus-stimulus associations in vmPFC neural populations, as we found representations for the predicted identity of subsequent cues when subjects saw a conditioned stimulus. Additionally, we found that vmPFC neural populations performed predictive value coding in consonance with amygdala populations. Taken together, these results suggest potential mechanisms through which model-based conditioned responses take place in the human brain.

Moreover, we found evidence that neurons in hippocampus, dACC and preSMA tracked the valence of outcome states containing unconditioned stimuli. Taking together single neuron encoding and population decoding results, activity in these brain areas also correlated with model-based estimates of how surprising it was to arrive at each outcome state, serving as potential substrates for a model-based learning signal.

Lastly, we investigated whether spiking activity in simultaneously recorded vmPFC and amygdala neuron pairs correlated in time, and found that cross-correlation patterns were modulated by the expected value of the conditioned stimulus. Specifically, amygdala neuron spikes tended to precede vmPFC spikes when patients were exposed to a cue which predicted a low valence outcome, while the opposite occurred for high value cues. This result suggests a mechanism for predictive value coding following learning through Pavlovian conditioning in humans.

Our findings provide support for a growing literature which suggests model-based learning as a mechanism for Pavlovian conditioning to take place (Dayan and Berridge, 2014). Previous work offered evidence implicating rodent OFC in model-based Pavlovian conditioning, following outcome revaluation (M. J. Robinson and Berridge, 2013) and sensory preconditioning (Jones et al., 2012; Sharpe, C. Y. Chang, et al., 2017). In humans, vmPFC/OFC were also shown to encode devaluation-sensitive predictive value coding, alongside with amygdala (Gottfried, O'Doherty, and Dolan, 2003), identity-specific representations of unconditioned stimuli (Klein-Flügge et al., 2013; Howard et al., 2015) and predictive stimulus-stimulus associations (Pauli, Gentile, et al., 2019), all of which are important aspects of creating cognitive maps for model-based learning.

The joint role of vmPFC/OFC and amygdala in predictive value coding can be under-

stood in a broader context, noting that these brain regions are deeply interconnected functionally and anatomically (Sharpe and Schoenbaum, 2016). For instance, primate OFC displays strong anatomical projections to amygdala (Aggleton, Burton, and Passingham, 1980), and vice-versa (Morecraft, Geula, and Mesulam, 1992), and the same is true for vmPFC proper (Joseph L Price, 1999). Importantly, neurons in OFC and amygdala have been shown to respond selectively in anticipation of rewarding or aversive outcomes (Schoenbaum, Chiba, and Gallagher, 1998; Kennerley and J. D. Wallis, 2009), and to the predictive value of cues in general (Padoa-Schioppa and Assad, 2006; Belova, Paton, and Salzman, 2008). Furthermore, lack of healthy amygdalar input induces significant changes to vmPFC activity during reward learning and decision making (Hampton, Adolphs, et al., 2007) and impairs the formation of neural ensembles in OFC to represent new contingencies during reversal learning (Schoenbaum, Setlow, et al., 2003). Our results are compatible with converging evidence positioning amygdala as a center for predictive value coding acting in consonance with OFC, which creates associations between stimulus identities and outcomes in the benefit of learning. We thus support this rich literature of interactions between vmPFC/OFC and amygdala and provide further mechanistic insight for value-based learning through the temporal correlations between spike trains in these brain areas.

Above and beyond its aforementioned role in predictive value coding in tandem with OFC, the amygdala has been shown to track expected values during decision making (Gottfried, O'Doherty, and Dolan, 2003; Holland and Gallagher, 2004; Hampton, Bossaerts, and O'Doherty, 2006; Salzman and Fusi, 2010; Wang, R. Yu, et al., 2017; Rudebeck, Ripple, et al., 2017), while lesion studies have suggested a causal role for this brain area in utilizing learned expected values for guiding behavior (Málková, Gaffan, and Murray, 1997; Bechara et al., 1999; De Martino, Camerer, and Adolphs, 2010). Additionally, the amygdala has been implicated in Pavlovian-instrumental transfer in humans (Talmi et al., 2008; Prévost, Liljeholm, et al., 2012) and in rodents, to the extent that targeted amygdala lesions in rodents have been shown to abolish motivational or identity-specific effects of Pavlovian cues on operant behavior (Corbit and Balleine, 2005).

Furthermore, we found widespread post-feedback encoding of outcomes and state prediction errors in cortex, especially in dACC and preSMA, in consonance with previous results implicating these regions in reward signaling (Kennerley, Behrens, and J. D. Wallis, 2011; Hill, Boorman, and Fried, 2016; Hunt et al., 2018) and model-

free reward prediction errors (Aquino et al., 2021). We also found above chance decoding of model-based EVs for the distal conditioned stimulus in preSMA, which we cautiously interpret as congruent with previous results found in human preSMA reporting integration of expected values in economic decision-making, in fMRI (Wunderlich, Rangel, and O'Doherty, 2009; Hare, Schultz, et al., 2011) and single neurons (Aquino et al., 2021).

Overall, our results provide evidence for model-based learning during Pavlovian conditioning in human vmPFC neurons, which encoded both stimulus-stimulus associations and expected values. Importantly, we found an effect of expected values over how amygdala and vmPFC neurons correlated in time, where vmPFC spikes tended to precede amygdala spikes when patients saw a cue predicting a rewarding outcome, and vice-versa for an unrewarding outcome. We also provide new evidence for how prefrontal cortex represents unconditioned stimuli and its effect on model-based learning following feedback. These findings shed light on single neuron representation of values and identity during the construction of cognitive maps in the context of Pavlovian conditioning and provide a general new perspective in how predictive value coding might generate conditioned responses.

3.5 Materials and Methods

Electrophysiology and recording

We used Behnke-Fried hybrid depth electrodes (AdTech Medical), positioned exclusively according to clinical criteria. Broadband extracellular recordings were performed with a sampling rate of 32 kHz and a bandpass of 0.1-9000Hz (ATLAS system, Neuralynx Inc.). The data set reported here was obtained bilaterally from hippocampus (HIP), amygdala (AMY), ventromedial prefrontal cortex (vmPFC), dorsal anterior cingulate cortex (dACC), and pre-supplementary motor area (preSMA) with one macroelectrode on each side. Each macroelectrode contained eight 40 μm microelectrodes. Recordings were bipolar, utilizing one microelectrode in each bundle of eight microelectrodes as a local reference.

Patients

Twelve patients (eight females) were implanted with depth electrodes for seizure monitoring prior to potential surgery for treatment of drug resistant epilepsy. One of the patients performed the task twice, totalling 13 recorded sessions. Human research experimental protocols were approved by the Institutional Review Boards of the California Institute of Technology and the Cedars-Sinai Medical Center.

Electrode location was determined based on preoperative and postoperative T1 scans obtained for each patient.

Pavlovian conditioning task

Patients performed a sequential Pavlovian conditioning task (Pauli, Larsen, et al., 2015) in which two fractals were presented in sequence, acting as distal and proximal conditioned stimuli (CSd, CSp, respectively). Following the CS pair, a video outcome was presented, either in the form of a hand depositing the patient's candy of choice into a paper bag, for a positive outcome, or in the form of an empty hand approaching a paper bag, for a neutral outcome. Patients were informed that each time they received a positive outcome contributed to a grand total of actual candy pieces they would receive at the end of the experiment. Patients chose among 5 possible candy options (Reese's Pieces, Hershey's Kisses, York Peppermint, Werther's Caramel, or Hi-Chew) to be delivered, according to their preference, and in the end of the session they received the closest possible equivalent to 200 calories in their candy of choice.

Trial structure is detailed in Fig. 3.1A. After a 1s fixation period, a distal CS was presented for 3s, followed by a 1s fixation period. Then, a proximal CS was displayed for 3s, followed by an outcome presentation video edited to a length of 3s. Intertrial intervals were jittered between 0.5s-1s. The distal CS was always presented along the vertical axis, either above or below the center of the screen, at a random position. Similarly, the proximal CS was always presented in a random position along the horizontal axis, inside one of two possible squares positioned to the left or the right of the screen. As an attention check, patients were asked to perform a button press whenever they saw a CSp inside one of the squares, reporting which of the two squares it was. To ensure this remained a purely Pavlovian paradigm, patients were instructed that these button presses did not affect the outcome of the trial in any way.

Each session contained a total of 96 trials, split in 4 blocks of 24 trials, with a two minute break between blocks. Within each block, there were 4 possible CS, 2 distal (A,B) and 2 proximal (X,Y). The identity of distal stimuli was re-selected in every block, but the two proximal stimuli were kept the same throughout the entire session, though their valences were reversed between blocks. There were four trial types in total (Fig. 3.1C): two common transitions (A-X-Reward and B-Y-No reward) and two rare transitions (A-Y-No reward and B-X-Reward), meaning that the CSp was

always fully predictive of the outcome, while the CSd was only partially predictive. Trial frequencies were 37.5% for each of the common types and 12.5% for each of the rare types. Every block always had exactly 9 trials of each common type and 3 trials of each rare type. Trial type order was selected randomly, except for the first 4 trials of each block, which we enforced to be of the common type, selected randomly.

The 10 fractals used in each session (2 CSd fractals for each block and 2 CSp fractals for the whole session) were chosen specifically for each patient out of a set of 24 possible fractals using the following procedure: before the beginning of the task, patients rated all 24 fractals for their subjective preference using a sliding bar from extremely unpleasant to extremely pleasant. The 10 stimuli which elicited the most neutral responses were selected for the experiment and assigned randomly to their CSd or CSp roles. For eye tracking, all the 24 fractals were to have the same luminance. In the end of every block, patients rated the 4 fractals they experienced for how pleasant they were. We z-scored ratings for each patient and used aggregate stimulus rating changes across patients as a candidate metric of Pavlovian conditioning.

Eye tracking

We tracked patients' pupil diameter during the task as a candidate conditioning metric (Pauli, Larsen, et al., 2015; Pauli, Gentile, et al., 2019; Pool et al., 2019), using a EyeLink 1000 camera (SR Research Ltd.) attached to the bottom of the task screen. We calibrated the camera using EyeLink's five point calibration before the session and between blocks. Pupil data was preprocessed to remove blinks and outlier points further than 5 s.d. from the mean diameter. We interpolated missing values removed in this way with the closest previous value and then filtered data with a 50ms moving average window. Pupil diameters were normalized in every trial relative to the initial 1s fixation period, using the average diameter in that period as a baseline. Statistical analyses were then performed using average diameter changes in the 0.5s-3s time windows after CSd and CSp presentations.

Computational model of learning

We used a normative model-based model to obtain estimates of how patients encoded transition probabilities between task states, stimulus expected values, and state prediction errors. We adapted a model used for a sequential instrumental task (Gläscher et al., 2010), to estimate a matrix $T(s, s')$ for the transition probabilities

from start states s to end states s'. Given distal states (A,B), proximal states (X,Y), and outcome states (reward, R; no reward, N), we defined the transition matrix T with transition probabilities $t_{ss'}$ as:

$$T = \begin{pmatrix} 0 & 0 & t_{AX} & t_{AY} & 0 & 0 \\ 0 & 0 & t_{BX} & t_{BY} & 0 & 0 \\ 0 & 0 & 0 & 0 & t_{XR} & t_{XN} \\ 0 & 0 & 0 & 0 & t_{YR} & t_{YN} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.1}$$

The t values were initialized to 0.5 and the remaining values in the matrix were chosen to be 0, reflecting the constraints of the task's transition structure. At each step, after transitioning from some state s to an end state s', a state prediction error is computed according to the following equation:

$$\delta_{SPE} = 1 - T(s, s') \tag{3.2}$$

The transition matrix T is then updated using a learning rate η, chosen normatively to be 0.22, adopting a value from a previosuly validated model-free model (Pauli, Larsen, et al., 2015).

$$T_{new}(s, s') = T(s, s') + \eta\delta_{SPE} \tag{3.3}$$

All the other values in T for states not transitioned into were also updated with $T_{new}(s, s'') = T(s, s'')(1 - \eta)$ to ensure each the sum of each row of T stays equal to 1, as it should reflect total probability.

Finally, the value of each state could be computed assuming the values of arriving into the reward and no reward states were 1 and 0, respectively:

$$EV_A = t_{AX}t_{XR} + t_{AY}t_{YR} \tag{3.4}$$

$$EV_B = t_{BX}t_{XR} + t_{BY}t_{YR} \tag{3.5}$$

$$EV_X = t_{XR} \tag{3.6}$$

$$EV_Y = t_{YR} \tag{3.7}$$

Since we were interested in tracking neural responses to stimulus identity associations, we used the inferred transition matrix T to determine the identity of the most likely proximal stimulus to follow distal stimulus presentation. We refer to this identity as *CSp presumed identity*.

Neural data pre-processing

We performed spike detection and sorting with the semiautomatic template-matching algorithm OSort (Rutishauser, Schuman, and Mamelak, 2006). Across all 13 sessions, we obtained 86 vmPFC, 165 amygdala, 119 hippocampus, 137 preSMA and 103 dACC putative single units (610 total). We refer to these isolated putative single units as "neuron" and "cell" interchangeably.

Single neuron encoding analysis

To quantify how single neuron activity correlated with several variables of interest, we performed a Poisson generalized linear model (GLM) analysis. As a dependent variable, we measured spike counts in three time windows: CSd window, from 0.25s to 3s after CSd presentation; CSp window, from 0.25s to 3s after CSp presentation; outcome window, from 0.25s to 2.25s after outcome presentation. Table 3.1 includes all dependent variables and the time windows in which they were tested.

Dependent variable	Time window
CSp presumed identity	CSd
CSp presumed EV	CSd
CSd EV	CSd
CSp identity	CSp
CSp EV	CSp
SPE	CSp
Outcome	Outcome
SPE	Outcome

Table 3.1: Dependent variables and respective time windows for Poisson GLM and population decoding analyses.

Population decoding analysis

We performed population decoding analyses by training a linear support vector machine (SVM) with MATLAB's function *fitcsvm*. In each session, we defined a population activity matrix X of dimensions $(nTrials, nNeurons)$ by counting

spikes in each trial within the same time windows from the encoding analysis. We then defined the decoded variable y as the same dependent variables from Table 3.1. To reduce the decoding problem to a classification task, we binned the continuous regressors (EV, SPE) into 3 tertiles before training the SVM with MATLAB's multiclass function *fitcecoc*. Cross-validation was performed by training on 2 trial blocks and testing on the 2 remaining trial blocks of each session. Since there were 4 blocks in total, we repeated the procedure for every possible combination of train/test blocks, resulting in 6 cross-validation folds. Test accuracies were averaged across folds and reported separately for each session. To obtain test accuracy significance levels, we repeated this entire procedure 500 times while shuffling the decoded variable y and compare the null mean test accuracy obtained in this manner with the true test accuracy, for each brain region. Finally, we corrected significance thresholds for the number of tested brain areas.

Cross-correlation analysis

To measure how neural activity across brain areas of interest was correlated, we performed a spike-spike cross-correlation analysis, with shuffle correction (Brody, 1999). First, we divided all trials in EV tertiles, excluding the middle tertile to obtain low EV and high EV trial groups. Then, for each trial group, we computed cross-correlations for each neuron pair containing neurons A and B recorded from the same session, but different brain areas (e.g. each neuron pair contained one vmPFC neuron vs. one amygdala neuron).

For two spike trains S_A^r, S_B^r (binned at 50ms with 5ms steps, constrained to CSp presentation time window), recorded from neurons A and B on trial r, we define the cross-correlogram of each trial as:

$$C^r(\tau) = \sum_{t=-\infty}^{\infty} S_A^r(t) S_B^r(t+\tau) := S_A^r \otimes S_B^r \tag{3.8}$$

Defining the <> operator as averaging across trials, we defined the shuffle-corrected cross-correlogram as:

$$V :=< (S_A^r - < S_A^r >) \otimes (S_B^r - < S_B^r >) > \tag{3.9}$$

$$V =< S_A^r \otimes S_B^r > - < S_A^r > \otimes < S_B^r > \tag{3.10}$$

The shuffle-correction procedure corrects for time-locked co-variation that might be caused to both neurons concurrently by stimulus presentation. Additionally, it ensures that the expected value of V is 0 if S_A and S_B are independent.

Finally, the reported shuffle-corrected correlation results were obtained by averaging the V vectors across all neuron pairs. To summarize the correlation results, we obtained the integrals from the positive and negative time lag regions and performed an ANOVA with EV level (low vs. high) and time lag sign (which region leads) as factors.

Chapter 4

VALUE-RELATED NEURONAL RESPONSES IN THE HUMAN AMYGDALA DURING OBSERVATIONAL LEARNING

The following chapter is adapted from Aquino et al., 2020 and modified according to the Caltech Thesis format.

Aquino, TG, Minxha, J, Dunne, S, Ross, IB, Mamelak, AN, Rutishauser, U, O'Doherty, JP, Value-related neuronal responses in the human amygdala during observational learning, Journal of Neuroscience (2020),

https://doi.org/10.1523/JNEUROSCI.2897-19.2020

4.1 Abstract

The amygdala plays an important role in many aspects of social-cognition and reward-learning. Here we aimed to determine whether human amygdala neurons are involved in the computations necessary to implement learning through observation. We performed single-neuron recordings from the amygdalae of human neurosurgical patients (male and female) while they learned about the value of stimuli through observing the outcomes experienced by another agent interacting with those stimuli. We used a detailed computational modeling approach to describe patients' behavior in the task. We found a significant proportion of amygdala neurons whose activity correlated with both expected rewards for oneself and others, and in tracking outcome values received by oneself or other agents. Additionally, a population decoding analysis suggests the presence of information for both observed and experiential outcomes in the amygdala. Encoding and decoding analyses suggested observational value coding in amygdala neurons occurred in a different subset of neurons than experiential value coding. Collectively, these findings support a key role for the human amygdala in the computations underlying the capacity for learning through observation.

4.2 Introduction

Acquiring new information about rewards associated with different stimuli is at the core of an animal's ability to adapt behavior to maximize future rewards (Sutton and Barto, 2018). In many organisms, reinforcement learning (RL) can take place

through taking actions and experiencing outcomes, but also indirectly through observing the actions taken and outcomes obtained by others, in a form of learning known as observational learning (OL) (Cooper et al., 2012; Van Den Bos, Jolles, and Homberg, 2013; Dunne, D'Souza, and O'Doherty, 2016; Charpentier and O'Doherty, 2018). The computational and neural basis of reinforcement-learning through direct experience has been the focus of intense study, and much is known about its neural underpinnings (Doya, 1999; Daw, Niv, and Dayan, 2005; D. Lee, Seo, and Jung, 2012; O'Doherty, Cockburn, and Pauli, 2017). In contrast, the neural mechanisms of observational learning have been much less well studied, especially in humans.

A core feature of RL models is that to decide whether or not to choose a particular stimulus or action, it is first necessary to consider the expected future reward associated with that option. Consistent with this, neuronal activity has been found in the amygdala as well as elsewhere in the brain, which tracks the expected future reward associated with various options at the time of decision-making (Gottfried, O'Doherty, and Dolan, 2003; Holland and Gallagher, 2004; Hampton, Bossaerts, and O'Doherty, 2006; Salzman and Fusi, 2010; Wang, R. Yu, et al., 2017; Rudebeck, Ripple, et al., 2017). Lesions of the amygdala have shown it is necessary for guiding behavior on the basis of expected future outcomes learned about through experience (Málková, Gaffan, and Murray, 1997; Bechara et al., 1999; Schoenbaum, Setlow, et al., 2003; Hampton, Adolphs, et al., 2007; De Martino, Camerer, and Adolphs, 2010), suggesting that value representations in this area are causally relevant for driving value-related behavior. The amygdala performs these functions in concert with a broader network that regulates reward learning, memory and emotion (Murray, 2007). Adaptive responses to reward cues and devaluation depend on amygdala-OFC connections in monkeys (Baxter et al., 2000) and mice (Lichtenberg et al., 2017). The amygdala also receives significant dopaminergic projections from the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) (Aggleton, Burton, and Passingham, 1980).

On the role of the amygdala in observational learning specifically, recent evidence has suggested a role for amygdala neurons in non-human primates in responding to the rewards obtained by others (S. W. Chang et al., 2015) as well as to others' choices (Grabenhorst, Báez-Mendoza, et al., 2019). However, much less is known about the role of the amygdala in the human brain in processes related to OL. One human single neuron study reported amygdala neurons which tracked observational

outcomes (Hill, Boorman, and Fried, 2016), though the most robust signals were found in rACC. Building on evidence implicating the human amygdala not only in reward processing but also in social cognition more broadly (Rutishauser, Mamelak, and Adolphs, 2015), we aimed to address how neurons in the human amygdala are involved in observational learning. We asked a group of neurosurgery patients to perform an observational learning task while we performed single-neuron recordings from electrodes in the amygdala. This provided us with a rare opportunity to investigate the role of amygdala neurons in value prediction coding and updating during OL. Since observational and experiential learning differ in that learning by observation is necessarily passive, to control for equivalence between observational and experiential learning, our task consisted of a passive Pavlovian paradigm, including instrumental trials exclusively to test contingency learning. We hypothesized we would find evidence for reinforcement-learning signals in the amygdala during observational learning, especially concerning the representation of the value of stimuli learned through observation. Furthermore, we also contrasted the contribution of amygdala neurons to OL with that of the role of these neurons in experiential learning. Of particular interest was the question of whether an overlapping or distinct population of neurons in the amygdala contributes to encoding reinforcement-learning variables in observational compared to experiential learning.

4.3 Materials and Methods

Electrophysiology and electrodes

Broadband extracellular recordings were filtered from 0.1Hz to 9kHz and sampled at 32 kHz (Neuralynx Inc). The data reported here was recorded bilaterally from the amygdala, with one macroelectrode in each side. Each of these macroelectrodes contained eight 40 μm microelectrodes. Recordings were performed bipolar, with one microwire in each area serving as a local reference (Minxha, Mamelak, and Rutishauser, 2018). Electrode locations were chosen exclusively according to clinical criteria.

Patients

Twelve patients (4 females) who were implanted with depth electrodes prior to possible surgical treatment of drug resistant localization related epilepsy volunteered to participate and gave informed consent. Four of the patients performed two recording sessions, and the others performed only one. One pilot session was not included in

the analysis and one session was discarded due to technical error. Protocols were approved by the Institutional Review Boards of the California Institute of Technology, the Cedars-Sinai Medical Center and the Huntington Memorial Hospital. Electrode location (Fig. 4.2d) was determined based on pre and post-operative T1 scans obtained for each patient. We registered each patients post-operative scan to their pre-operative scan, which we in turn registered to the CIT168 template brain (Tyszka and Pauli, 2016) (which is in MNI152 coordinates) using previously published methods (Minxha, C. Mosher, et al., 2017).

Electrode localization, spike detection and sorting

Spike detection and sorting was performed as previously described using the semiautomatic template-matching algorithm OSort (Rutishauser, Schuman, and Mamelak, 2006). Channels with interictal epileptic activity were excluded. Across all valid sessions, we isolated in total 202 putative single units in amygdala (135 in right amygdala (RA) and 67 in left amygdala (LA)). We will refer to these putative single units as "neuron" and "cell" interchangeably. Units isolated from electrodes localized outside of the amygdala were not included in the analyses. Using the probabilistic CIT168 atlas of amygdala nuclei (Tyszka and Pauli, 2016), we determined the subnuclei from which the unit was recorded from: deep or basolateral (BL), with 117 units; superficial or corticomedial (CM), with 39 units; and remaining nuclei (R), with 46 units, which contained neurons from either the anterior amygdaloid area or the central nucleus (we were not able to distinguish between the two). We characterized the quality of the isolated units using the following metrics: the percentage of interspike intervals (ISIs) below 3 ms was 0.49% ± 0.63%; the mean firing rate was $1.98Hz \pm 2.47Hz$; the SNR at the mean waveform peak, across neurons, was 5.12 ± 3.24; the SNR of the mean waveform across neurons was 1.87 ± 0.97; the modified coefficient of variation (CV2) (Holt et al., 1996) was 0.95 ± 0.11; and the isolation distance (Schmitzer-Torbert et al., 2005) was 1.69 ± 0.59 for neurons in which it was defined.

Task and behavior

Patients performed a multi-armed bandit task (Dunne, D'Souza, and O'Doherty, 2016) with 288 trials in total, distributed across 2 experiential and 2 observational blocks. Block order was chosen to always interleave block types, and the type of the initial block was chosen randomly (see Fig. 4.1a). Each block had 72 trials, out of which 48 were no-choice trials and 24 were binary choice trials. Choice

and no-choice trials were randomly distributed across each block. Experiential no-choice trials began with the presentation of a single bandit, whose lever was pulled automatically 0.5s after stimulus onset. Each block consisted of two possible bandits that were chosen randomly in every trial. Subjects were told that the color of a bandit allows them to differentiate between the different bandits. Bandits were repeated across blocks of the same type, with the possibility of contingency reversal. Reversals could happen only once in the entire task. For the sessions that did include a reversal (9 out of the 14 analyzed sessions), it always happened right before the beginning of the third block of trials, at the halfway point in the session. Patients were not told in advance about the reversals, but were fully instructed about the reward structure of the task (as explained below).

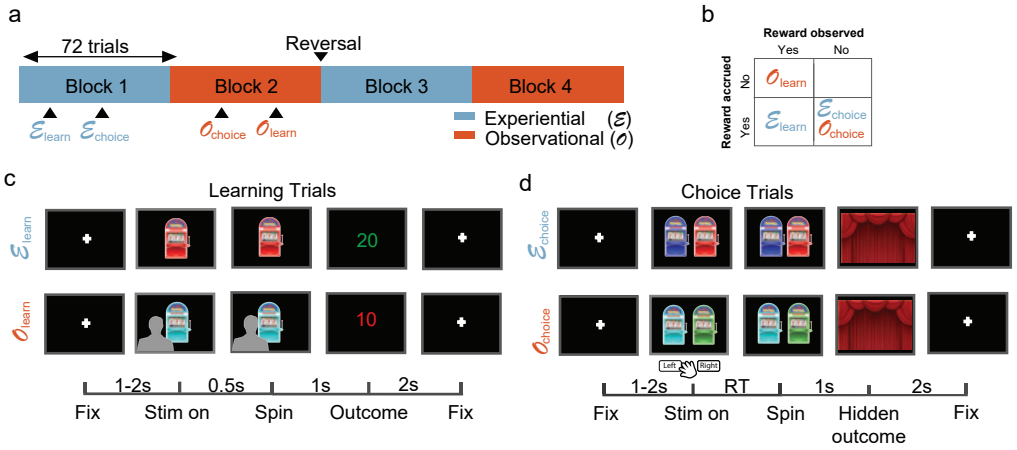

Figure 4.1: Observational learning task. (a) Block structure. The task had 288 trials in total, in 4 blocks of 72 trials. Each block contained either experiential or observational learning trials, as well as choice trials. Block order was interleaved, and bandit values were reversed after the end of block 2. (b) Reward structure. Reward was accrued to subjects' total only in experiential trials, and reward feedback was only presented in learning trials, both in experiential and observational blocks. (c) Learning trials structure. Top row: experiential learning trials. After a fixation cross of jittered duration between 1-2s, subjects viewed an one-armed bandit whose tumbler was spun after 0.5s. After a 1s spinning animation, subjects received outcome feedback, which lasted for 2s. Bottom row: observational learning trials. Subjects observed a video of another player experiencing learning trials with the same structure. Critically, outcomes received by the other player were not added to the subject's total. Lower bar: timing of trial events in seconds. (d) Choice trials structure. Subjects chose between the two bandits shown in the learning trials of the current block. After deciding, the chosen bandit's tumbler spun for 1s, and no outcome feedback was presented.

Outcome was presented 1s after the automatic lever press, and subjects received

feedback on the amount of points won or lost in the trial, which was added or subtracted to their personal total. The amount of points for each trial was selected from a normal distribution, with specific means and variances for each bandit, truncating at -50 and +50 points. Subjects could not see their added points total during the task, but were shown their overall points total after the end of the task.

Observational no-choice trials consisted in watching a pre-recorded video of another player experiencing the same trial structure. These videos contained the back of the head of a person (always the same individual), as he watched a screen containing a bandit playing out a no-choice trial, including the outcome in points. The person did not move during the video, but the bandit on the screen was animated to display the lever press and outcome display, same as in experiential trials.

Points received by the other player in the pre-recorded video were not added or subtracted to subjects' personal total, and subjects were informed of this fact. Choice trials started with the presentation of two bandits, and subjects had up to 20s to select one via button press, which would cause the lever on the corresponding bandit to be pulled. If subjects failed to respond within 20s, the trial was considered missed, and subjects received a penalty of 20 points. In choice trials, after a 1s period, subjects observed closed curtains on the screen instead of outcome feedback. Subjects were told they should attempt to maximize the amount of received rewards, and that they would still receive or lose the amount of points displayed behind the curtain, despite the lack of feedback. This was done to restrict learning to no-choice trials, to further dissociate the decision making and reward learning components of the task. The two bandits that could be chosen in choice trials were always the two possible bandits from no-choice trials in the current block. Intertrial intervals were jittered with a uniform distribution between 1s and 3s regardless of block and trial type. To further motivate subjects, a leaderboard was shown in the end of the task, displaying the amount of points won by the subject, in comparison to amounts won by previous participants.

Computational modeling

We focused on the form of observational learning referred to as vicarious learning, which takes place when individuals observe others taking actions and experiencing outcomes, rather than doing so themselves (Charpentier and O'Doherty, 2018). At the computational level, we hypothesized that vicarious learning involves similar mechanisms to those utilized for experiential learning. To test this hypothesis, we

adapted a simple model-free learning algorithm from the experiential to the observational domain (Cooper et al., 2012). For both observational and experiential learning, this model learns EVs for each stimulus via a reward prediction error (RPE) that quantifies discrepancies between predicted values and experienced reward outcomes. This prediction error signal is then used to iteratively update the value predictions.

We used the behavior in the choice trials to fit four different types of computational models. We used a hierarchical Bayesian inference framework to achieve both hierarchical model fitting and model comparison (Piray et al., 2019). This framework allowed us to infer a protected exceedance probability for each model, as well as individualized model parameters for each subject. The model with the largest protected exceedance probability was chosen for the model-based encoding and decoding analyses. The exceedance probability value expresses the probability that each model is the most frequent in the comparison set (Rigoux et al., 2014). Protected exceedance probability is a typically more conservative metric which takes into account the possibility that none of the compared models is supported by the data (Piray et al., 2019).

We first provide a brief summary of each of the computational models before describing each in detail. The first model was a simple reinforcement learning model (Sutton and Barto, 2018) with a single learning rate parameter for both experiential and observational trials (RL (no split)); the second model was the same, except that learning rates were split between observational and experiential trials (RL (split)); the third model was a counterfactual reinforcement learning model with a single learning rate in which EVs for played bandits were updated as usual, but EVs for the bandits that were not seen in a trial were also updated, in the opposite direction of the bandits that were actually played (RL (counterfactual)). The last model was a hidden Markov model with built-in reversals, with two states. The first state assumed one of the bandits in the block had a positive mean payout, while the other bandit had a negative mean payout with the same magnitude. The second state mirrored the first one, switching which bandits had the positive and negative payouts. This model allowed us to include inferred reversals between those two states, and to model the inferred reversal rate that patients assumed to be true. Expected values in all models were initialized to zero for all bandits.

The RL (no split) model keeps a cached value V for the EV of each bandit i, in every trial t, updated according to the following rule:

$$V_i^{(t+1)} = V_i^t + \alpha\delta_t \tag{4.1}$$

$$\delta_t = R_t - V_i^t \tag{4.2}$$

In this case, α represents the learning rate for both the experiential and observational cases, δ represents reward prediction error (RPE) and R represents reward feedback value. The RL (split) model is identical, except that a learning rate α_{exp} is applied in experiential trials and another learning rate α_{obs} is applied in observational trials.

The RL (counterfactual) model is identical to RL (no split), except that both bandits are updated on every trial, in opposite directions. For the chosen and unchosen bandits in every trial, the EV updates are as follows:

$$V_{chosen}^{(t+1)} = V_{chosen}^t + \alpha\delta^t \tag{4.3}$$

$$V_{unchosen}^{(t+1)} = V_{unchosen}^t - \alpha\delta^t \tag{4.4}$$

The HMM has been formalized similarly to previous work (Prévost, McNamee, et al., 2013). An inferred state variable S_t represented the association between bandits and rewards at trial t. Assuming the two bandits in a block are arbitrarily indexed as A and B, that the magnitude of the inferred mean payout was a free parameter μ fixed throughout the task, and that mu_A and mu_B denote the mean payouts for bandits A and B respectively,

$$S_t = \begin{cases} 0, & \text{if } \mu_A = +\mu, \mu_B = -\mu \\ 1, & \text{if } \mu_A = -\mu, \mu_B = +\mu \end{cases} \tag{4.5}$$

This model allows for inferring reversals between states, which means the inferred mean payouts of the two bandits are swapped. The reversal structure is dictated by the following reversal matrix, assuming reversal rates were a free parameter r fixed throughout the task:

$$P(S_t|S_{t-1}) = \begin{pmatrix} 1-r & r \\ r & 1-r \end{pmatrix} \tag{4.6}$$

Given this transition structure, the prior $Prior(S_t)$ would be updated in every trial as follows:

$$Prior(S_t) = \sum_{S_t states} P(S_t|S_{t-1})P(S_{t-1}) \tag{4.7}$$

Initial state probabilities were set to 0.5. Then, using Bayes' rule, the posterior would be updated using evidence from the outcome R_t:

$$PosteriorP(S_t) = \frac{P(R_t|S_t)Prior(S_t)}{\sum_{S_t states} P(R_t|S_t)Prior(S_t)} \tag{4.8}$$

Outcome variables were assumed to have a Gaussian distribution, with a fixed standard deviation free parameter σ:

$$R_t|(S_t = 0) \sim \begin{cases} N(+\mu, \sigma), & \text{for bandit A} \\ N(-\mu, \sigma), & \text{for bandit B} \end{cases} \tag{4.9}$$

$$R_t|(S_t = 1) \sim \begin{cases} N(-\mu, \sigma), & \text{for bandit A} \\ N(+\mu, \sigma), & \text{for bandit B} \end{cases} \tag{4.10}$$

This framework allowed for computing EVs in each trial t for each bandit j, taking into account the probability of being in each state:

$$\begin{aligned} EV_j(t) = E[R|\text{ bandit j, trial t}] = \\ P(S_t = 0) \int_{Outcomes} R\ P(R|S_t = 0,\ \text{bandit j}) dR + \\ P(S_t = 1) \int_{Outcomes} R\ P(R|S_t = 1,\ \text{bandit j}) dR \end{aligned} \tag{4.11}$$

Since outcomes were assumed to be normally distributed, for each bandit this reduced to

$$EV_A(t) = P(S_t = 0)\ \mu - P(S_t = 1)\ \mu = \mu(P(S_t = 0) - P(S_t = 1)) \tag{4.12}$$

$$EV_B(t) = -P(S_t = 0)\ \mu + P(S_t = 1)\ \mu = \mu(P(S_t = 1) - P(S_t = 0)) \tag{4.13}$$

This means that EVs for a certain bandit were larger if patients inferred they were more likely in the state in which that bandit was better. For example, if $P(S_t = 0) = 0.9$ and $P(S_t = 1) = 0.1$, then:

$$EV_A = \mu(0.9 - 0.1) = 0.8\mu \tag{4.14}$$

$$EV_B = \mu(0.1 - 0.9) = -0.8\mu \tag{4.15}$$

We used the cached value estimates of EV as parameters in a softmax function controlled by an inverse-temperature parameter β for each session, to generate decision probabilities in free-choice trials. For the RL models, we constrained α, α_{exp}, and α_{obs} in the (0,1) interval, and β in the (0,10) interval. In the HMM, we constrained r in the (0,1) interval, and both μ and σ in the (0,20) interval.

Model comparison was performed by computing the protected exceedance probability of each model and selecting the one with the largest value, which was RL (counterfactual) (Fig. 4.3). For all subsequent model-based analyses, we display results using the EVs and RPEs produced by the RL (counterfactual) model. An example of how the EVs assigned to each bandit typically behave in a modeled session is displayed in Fig. 4.2b.

Population decoding analysis

Population decoding was performed with the Neural Decoding Toolbox (Meyers, 2013) as described previously (Rutishauser, Ye, et al., 2015). We pooled neurons from all sessions into a single pseudopopulation with 202 amygdala neurons. To achieve this alignment on a trial by trial basis across sessions, we created discrete trial bins using quantiles of the decoded variable, with the same number of trials, for each session. For example, for outcome decoding, we found which trials for each session fit into each one of 4 quantiles of received outcomes and aligned trials that fell in the same bin across sessions, assuming all neurons belonged to the same session. This session-based trial binning meant the exact quantile boundaries were not necessarily the same across sessions. For example, a trial in which the outcome was 10 points might have been placed on bin 2 for one session and on bin 3 for another session, depending on the distribution of outcomes in each session. Finally, all learning trials across sessions had the same event timing, so no additional temporal alignment was needed.

We used this strategy to create a neural activity tensor of dimensions $(n_{neurons}, n_{trials})$, where n_{trials} is the number of trials in a single session. Decoding consisted of training and testing a classifier tasked with correctly predicting which bin of the variable of interest each trial belonged to, only from information contained in the neural activity tensor.

We used a maximum Pearson correlation classifier with access to spike counts binned in a time window of interest. This classifier learns by obtaining a mean representation x_c of each class c in the multidimensional neural population space, and assigns new data points y to a class c^* corresponding to $c^* = argmax_c(corr(x_c, y))$. We used 10-fold cross-validation and 20 cross-validation fold resample runs, which were averaged to generate a testing decoding accuracy score. Significance was determined via permutation test with 500 re-runs, shuffling regressor labels. Expected value decoding was only tested in the pre-outcome period (300ms to 1500ms from trial onset), whereas outcome and prediction error decoding was only tested in the post-outcome period (300 to 2000ms from outcome onset).

Single neuron encoding analysis

For every tested neuron n, we used a Kruskal-Wallis test (Kruskal and W. A. Wallis, 1952) to fit binned spike counts $y_n(t)$ (1200 ms bins for pre-outcome, 1700 ms bins for post-outcome, 3500ms bins for the whole trial), implemented with the MATLAB function *kruskalwallis*. Outcome and prediction errors were regressed only on the post-outcome period, and expected values were regressed only on the pre-outcome period. Trial type regression was performed in the entire trial. Significance was determined through permutation tests by shuffling variable labels. For expected value and prediction error time series, in which trials might not be independent from each other, we performed variable shuffling using surrogate time series as described previously (Schreiber and Schmitz, 2000). For the other variables, we used standard random permutations. We then used chi-squares yielded by the Kruskal-Wallis test as a statistic for each regressor.

Simulation for comparing encoding and decoding

In order to better understand possible discrepancies between the encoding and decoding analyses, we set out to simulate a population of artificial neurons responding to a categorical variable, and to compare encoding and decoding analyses within this population for varying levels of noise. Each simulation contained a population of 200 Poisson point processes as artificial neurons in 96 trials. We created an

artificial categorical variable X to be encoded and decoded, whose value would be sampled randomly with a uniform distribution in each trial from $[1, 2, 3, 4]$. The latent firing rate λ_n of each neuron n was then given by:

$$\lambda_n = \exp(\mu_n X + \epsilon) \tag{4.16}$$

The factor μ_n scales the influence of the categorical variable X over latent firing rates. For every neuron n, and for every simulation, we sampled μ_n from a normal distribution $N(0, 0.4)$. The factor ϵ controls the amount of random noise added to latent firing rates. For every trial, ϵ was sampled from a normal distribution $N(0, 0.4\sigma)$, where σ is the noise factor variable. We ran 100 simulations with each one of the following noise factors σ: $[1, 5, 20]$. We included 0.4 as a factor in the distributions of μ_n and σ only with the intent of generating plausible spike counts.

Following the construction of latent firing rates, we simulated how many spikes occurred in a time window of 1 second, and used these spike counts as the input for encoding and decoding analyses. We performed encoding by applying the previously described Kruskal-Wallis test, using the artificial spike counts and categories. We obtained chance levels for the encoding analysis theoretically, from an inverse binomial distribution, assuming a chance level of 0.05 and a total neuron population of 200. Additionally, we performed decoding of the variable X from spike counts with the previously described maximum Pearson correlation classifier, with a $75-25\%$ split between training and testing trials, re-sampling cross-validation folds 10 times. Chance levels in decoding single test trials were obtained theoretically from an inverse binomial distribution, assuming a chance level of 0.25 (since X had 4 categories).

Finally, we compared (Fig. 4.5h) how well decoding and encoding analyses performed for varying levels of the noise factor σ, in terms of decoding accuracy in single test trials, as well as significant neuron count.

4.4 Results

Behavioral performance

We obtained a behavioral metric of subject performance on choice trials (Fig. 4.2a): we defined "correct" trials as those in which the subject selected the bandit with the highest mean payout, disregarding the first 25% of trials in each block. The reason we excluded the first 25% of trials from accuracy analysis only was to get a coarse metric of overall accuracy discarding the transient initial period of learning. Note

this transient period is still of interest in terms of measuring expected values and reward prediction errors, so it is still included in the computational modeling and the neural analysis. We found that in 12 out of 15 recording sessions, performance was above the 95% percentile of a random agent, theoretically determined by a binomial distribution with success probability 0.5, thereby indicating that behavior in these sessions was significantly better than chance.

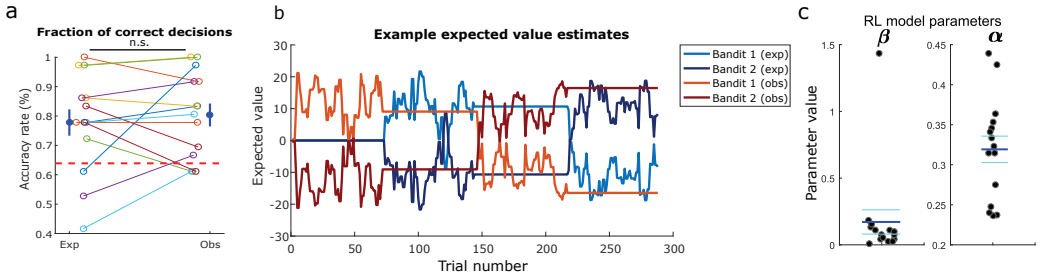

Figure 4.2: Behavior and reinforcement learning model. (a) Accuracy rate of all sessions as defined by the fraction of free trials in which a subject chose the bandit with highest mean payout, discarding the first 25% of trials in each block. Each color represents a different session, for experiential and observational trials, with average and standard error indicated on the left and right. Accuracy in experiential and observational trials was not significantly different ($p < 0.66$, two-sample t-test). The dashed red line indicates the chance level estimated by the theoretical 95% percentile of correct proportions, obtained from an agent making random decisions with $p = 0.5$. (b) Typical time course of modeled EVs throughout the task, using the RL (counterfactual) model. Bandit 1 (exp) and Bandit 2 (exp) indicate EVs for each of the two bandits shown in experiential blocks respectively, whereas Bandit 1 (obs) and Bandit 2 (obs) indicate EVs for each of the two bandits shown in observational blocks, respectively. (c) Parameter fits for each valid session, for the chosen reinforcement learning model. The model contained a single learning rate (α)for experiential and observational trials, and an inverse temperature β. Dark blue horizontal lines indicate parameter means and cyan horizontal lines indicate standard error.

Overall, subjects performed well in both the experiential and observational condition (Fig. 4.2a): The proportion correct in all trials was 0.776 ± 0.038. Taking only experiential trials, the proportion correct was 0.763 ± 0.044; in observational trials, it was 0.789 ± 0.039. Experiential and observational correct proportions were not significantly different from each other across all sessions (two sample t-test, $p = 0.6641 > 0.05$). We estimated the chance level using the theoretical 95% percentile of correct proportions, obtained from an agent making random decisions with $p = 0.5$, assuming a binomial distribution. We also tested whether the correct

proportions were different, in all the trials following a reversal, for the patients that did experience a reversal, between observed (0.78 ± 0.072) and experiential trials (0.800 ± 0.074), but found no evidence of a difference (two-sample t-test, two-tailed, $p = 0.894$).

Computational model fitting

We fit four computational models to each subjects' behavior during choice trials (see methods): a model-free reinforcement learning model with one learning rate for experiential and observational trials (RL (no split)); a model-free reinforcement learning model with separate learning rates split between experiential and observational trials (RL (split)); a counterfactual reinforcement learning model in which outcomes from the played bandit also were used to update EVs for the unseen bandit in each trial; and a hidden Markov model (HMM) with an estimate of reversal rates on a trial-by-trial basis. In all models, we applied a softmax rule to generate probabilistic decisions. Model fitting and comparison were performed simultaneously with hierarchical Bayesian inference (HBI) (Piray et al., 2019), described in more detail in the methods section.

Overall, the counterfactual RL model, with a single learning rate for experiential and observational trials, outperformed the others in both protected exceedance probability (Fig. 4.3a) and inferred model frequency among the patient population (Fig. 4.3b). The mean learning rate in the winning model was 0.31 ± 0.06, and the mean softmax inverse temperature β was 0.17 ± 0.35 (Fig. 4.2c). Using HBI, we also compared the single learning rate counterfactual model with a counterfactual model which had split learning rates between experiential and observational trials, finding a 99.9% protected exceedance probability for the single learning rate counterfactual model. Taken together, these behavioral findings suggest subjects employed a similar learning strategy for the valuation of each bandit regardless of trial type, and were still engaged with the task when another person received rewards. Given that the counterfactual model was the best fitting model for explaining participants' behavior on the task, we utilized the variables generated by this model in the subsequent computational model-based analysis of the neuronal data.

Amygdala population decoding

We tested whether the activity of amygdala neurons was related to the following task and computational variables: trial type, EV, outcome, and RPE, during learning trials. Trial type decoding was performed in the whole trial; EV decoding was

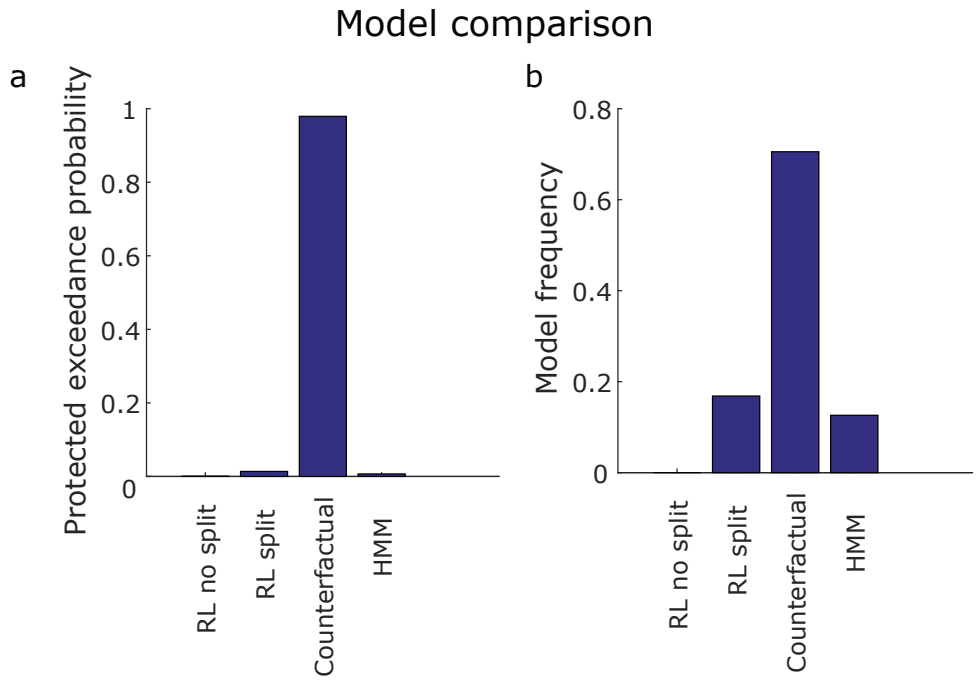

Figure 4.3: Model comparison. (a) Protected exceedance probability. This is the probability that each one of the four models fit using hierarchical Bayesian inference (RL split, RL no split, Counterfactual, and HMM) was more likely than any other, taking into account the possibility that there is no difference between models. (b) Model frequency. This is the proportion of individual patients whose behavior is better explained by each model. The counterfactual learning model outperforms the others both in terms of protected exceedance probability and model frequency.

performed in a 1200ms time bin starting 300 ms from stimulus onset, until outcome presentation; outcome and RPE decoding was performed in a 1700ms time bin starting 300ms after outcome presentation.

For each variable, we trained a maximum Pearson correlation classifier on a pseudopopulation of amygdala neurons (see methods; Fig. 4.4). Cross-validated single-trial decoding accuracy was obtained for each tested variable, tested for significance through a permutation test with 500 shuffled label runs. The same procedure was repeated in 50 cross-validation randomly re-sampled folds. To perform decoding of continuous variables across sessions, we binned variables (EV, outcome, and RPE) into 4 bins (quantiles). P-values were obtained by computing the proportion of shuffled instances in which decoding accuracy exceeded the real decoding accuracy. With this method, the smallest p-value attainable was $1/n_{permutations} = 0.002$.

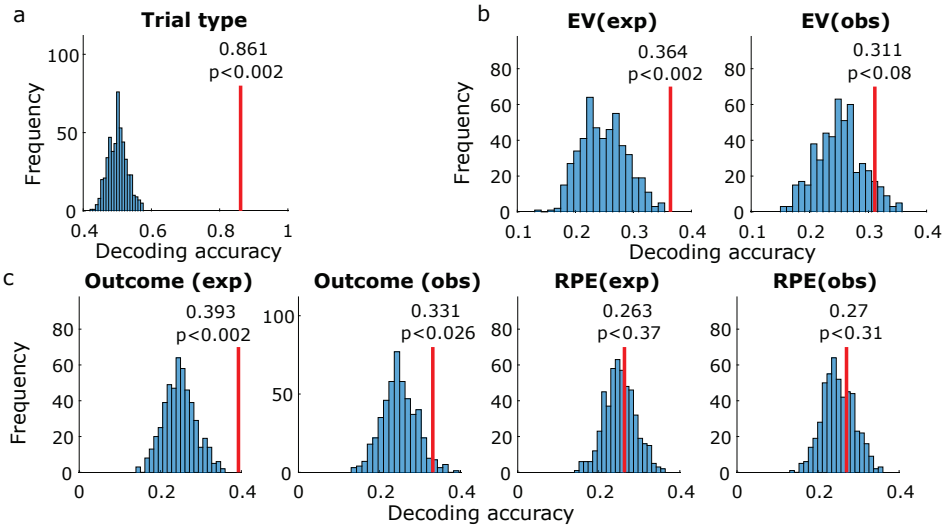

Figure 4.4: Amygdala population decoding analysis. (a) Entire trial decoding. The tested variable was trial type (experiential vs. observational). The vertical red line indicates average decoding accuracy in held-out trials after training with a maximum Pearson correlation classifier. The histogram indicates decoding accuracy in each instance of a permutation test, shuffling variable labels. P-values were obtained by computing the proportion of permutation iterations in which the decoding accuracy exceeded the true decoding accuracy. (b) Same, decoding within the pre-outcome period. Decoded variables, from left to right, were EV (experiential) and EV (observational). (c) Same, decoding within the post-outcome period. Decoded variables, from left to right, were outcome (experiential), outcome (observational), RPE (experiential), and RPE (observational).

Trial type decoding

Trial type (experiential vs observational) could be decoded from amygdala neurons with above chance accuracy (see Fig. 4.4a; $p < 0.002 < 0.05$, permutation test). Average decoding accuracy in held-out trials was 86.1%. This indicates that amygdala neurons prominently tracked whether the current block was experiential or observational.

Expected value decoding

We next tested whether EV was decodable in the pre-outcome period (300ms to 1500ms from bandit onset), separately for observational and experiential learning trials. We found better than chance decoding in experiential trials (see Fig. 4.4b; $p < 0.002 < 0.05$, permutation test). Average experiential EV decoding accuracy in the pre-outcome period was 36.4%. In contrast, observational EV decoding was

within the chance boundaries of the permutation test ($p < 0.08$). (Fig. 4.4b). This indicates that amygdala neuron populations contained more easily decodable information for keeping track of rewards received by oneself than by the other player.

Outcome decoding

Following outcome onset (300ms to 2000ms from outcome onset), outcome was decodable above chance in experiential trials (see Fig. 4.4c; $p < 0.002 < 0.05$, permutation test), with an average decoding accuracy of 39.3%. Additionally, outcome was also decodable above chance in observational trials (see Fig. 4.4c; $p < 0.026 < 0.05$, permutation test), with an average decoding accuracy of 33.1%. This indicates that amygdala populations represented both experienced and observed outcomes, but more strongly in the experienced case.

Reward prediction error (RPE) decoding

We tested for decodability of RPEs during the outcome period (300ms to 2000ms from outcome onset), but did not find better decoding accuracy than expected by chance in the permutation test (see Fig. 4.4c), both in the experiential ($p < 0.37$, permutation test) and the observational cases ($p < 0.31$, permutation test).

Single neuron encoding analysis

In order to understand the relationship between the population decoding result and the activity of single neurons we next tested the sensitivity of each amygdala neuron ($n = 202$ neurons) to each one of the decoded variables (Fig. 4.5). We used a Kruskal-Wallis analysis to compare every individual neuron's activity to the same variables used in decoding. We chose this method as opposed to a GLM analysis to encompass units whose activities might be non-linearly modulated by a variable of interest (e.g. being less active for intermediate levels of a variable of interest), such as the one displayed in (Fig. 4.6d).

Trial type neurons

We found 100 amygdala neurons whose activity is significantly different across experiential and observational trials (49.5%, $p < 0.002 < 0.05$, permutation test). Note this could partially be explained as an effect of the blocked design we chose, grouping all experiential trials and observational trials in distinct trial blocks. This

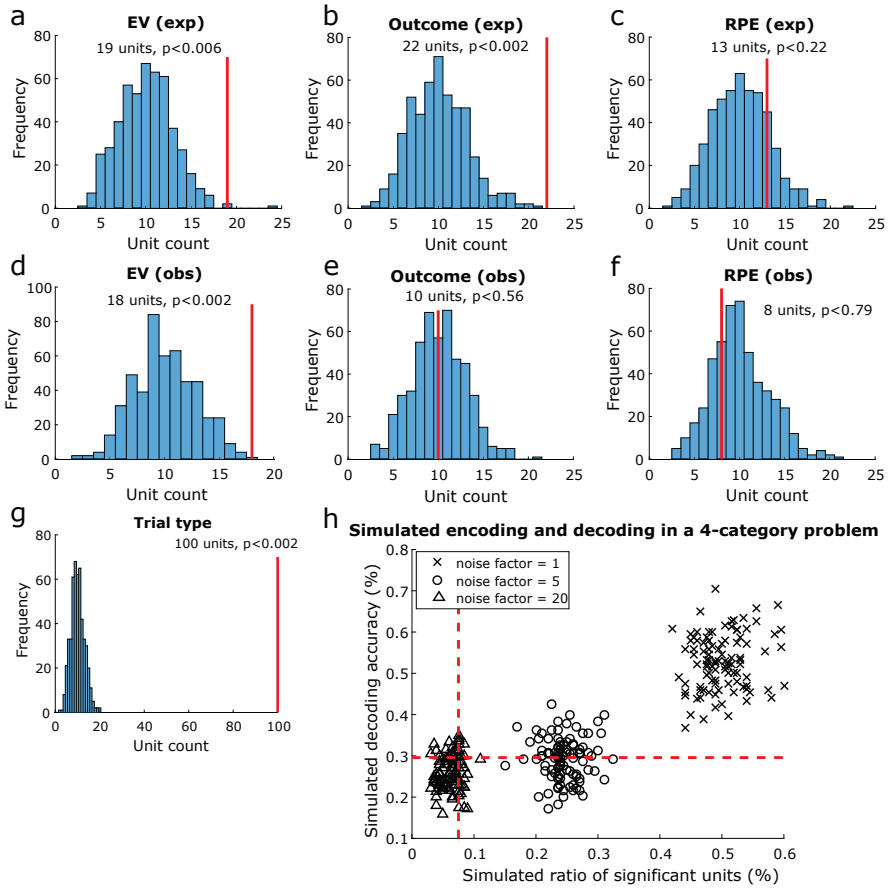

Figure 4.5: Amygdala single neuron encoding analysis. (a) Pre-outcome encoding of EV in experiential trials. Solid red lines indicate how many units were found to be sensitive to the tested variable within the pre-outcome period. Histograms sensitive units in the permutation test. Similarly, we tested encoding for (b) experiential outcome in the post-outcome period; (c) experiential RPE in the post-outcome period; (d) observational EV in the pre-outcome period; (e) observational outcome in the post-outcome period; (f) observational RPE in the post-outcome period; (g) trial type in whole trials. (h) Comparing encoding and decoding in a simulated 4-category problem with varying noise levels. Noise was added to the latent firing rate of each neuron scaled by a noise factor of 1 (crosses); 5 (circles); or 20 (triangles), and simulations were repeated 100 times for each noise level. Each data point in the plot represents one individual simulation. Dashed red lines indicate theoretical chance levels for encoding (vertical) and decoding (horizontal).

result is also consistent with the high trial type decoding accuracy we found in left-out trials.

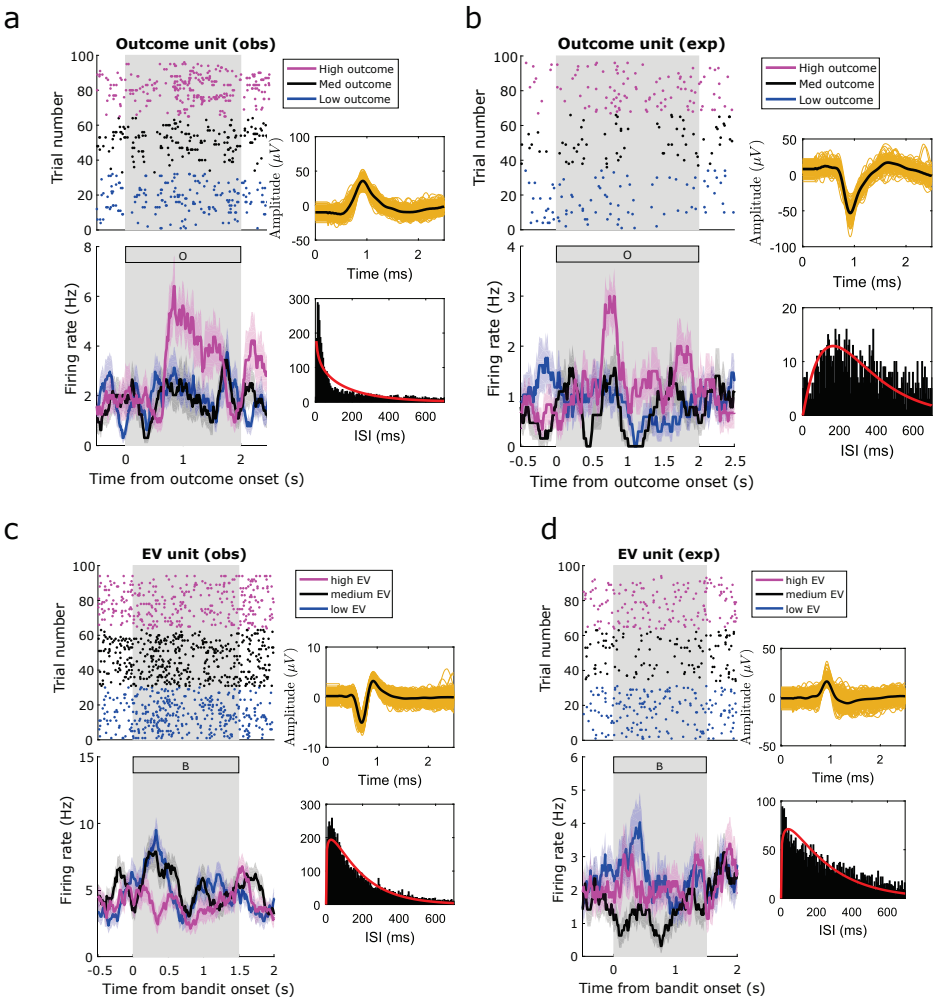

Figure 4.6: Amygdala neuron raster plot examples. Example amygdala units, significantly modulated by the indicated regressors, in the indicated conditions. (a) Unit modulated by outcome in observational trials during post-outcome period. (b) Unit modulated by outcome in experiential trials during post-outcome period. (c) Unit modulated by EV in observational trials during pre-outcome period. (d) Unit modulated by EV in experiential trials during pre-outcome period. Top: raster plots. For plotting purposes only, we reordered trials by regressor levels by obtaining 3 quantiles from the variable of interest (magenta: high; black: medium; blue: low). Bottom: PSTH (bin size = $0.2s$, step size = $0.0625s$). The annex panels to the right of each raster display spike waveforms (top) and interspike interval histograms (bottom) from the plotted neuron. Background gray rectangles post-outcome periods, for (a) and (b), or pre-outcome periods, for (c) and (d). Rectangles filled with a letter indicate which stimulus was present on the screen at that time (B: bandit; O: outcome).

Expected value neurons

We tested amygdala neurons for experiential EV sensitivity, and found 19 sensitive units (9.4%, $p < 0.006 < 0.05$, permutation test) during the pre-outcome period

(Fig. 4.5b, left). One experiential EV example unit is shown in Fig. 4.5d. Conversely, observational EV sensitivity was found in 18 units (8.9%, $p < 0.002 < 0.05$, permutation test). An observational EV example unit is shown in Fig. 4.5c. Taken together, these findings suggest that the expectation of outcomes is represented in a significant proportion of amygdala neurons, both for experienced and observed outcomes.

Outcome neurons

We also tested amygdala neurons for outcome sensitivity, in the post-outcome period. For experiential outcomes, we found a significant proportion (10.8%, $p < 0.002 < 0.05$, permutation test) of sensitive amygdala neurons (Fig. 4.5c, first panel). For observational outcomes (Fig. 4.5c, second panel), however, only 10 units were selected as sensitive (4.9%, $p < 0.56$, permutation test), despite better-than-chance observational outcome decoding. Example outcome neurons are displayed in Fig. 4.6a (observational) and Fig. 4.6b (experiential).

Reward prediction error neurons

Also in the post outcome period, we found 13 (6.4%, $p < 0.22$, permutation test) experiential RPE units (Fig. 4.5c, third panel), as well as 8 (3.9%, $p < 0.76$, permutation test) observational RPE units (Fig. 4.5c, fourth panel). Neither of these unit counts exceeded what is expected by chance in the permutation test. This finding is consistent with the low decoding accuracy we obtained for reward prediction errors in the population decoding analyses.

Anatomical location

We used a chi-squared test of independence (1 degree of freedom) to determine whether units located in the right or left amygdala were more likely to be sensitive to each variable tested (Table 4.1). We found no evidence of lateralization for any of the tested variables. Similarly, we used a chi-squared test of independence (2 degrees of freedom) to test whether units were more likely to be sensitive to each variable in some amygdalar subnuclei group, the null hypothesis being that all groups were equally likely to contain units sensitive to each tested variables (Table 4.2). We found no evidence of any group being more likely than the others to contain sensitive units for any variable. Taken together, these findings provide no evidence for spatial specialization of value-related variables in amygdala, either

by lateralization or by specialization within subnuclei. These findings must be interpreted cautiously, however, given our relatively low unit counts in each tested subset of amygdala neurons.

	Right (n=135)	Left (n=67)	p-value
Trial type	64	36	0.397
EV (exp)	11	8	0.384
EV (obs)	13	5	0.611
Outcome (exp)	14	8	0.736
Outcome (obs)	7	3	0.827
RPE (exp)	8	5	0.675
RPE (obs)	3	5	0.072

Table 4.1: Sensitive units by side. Number of sensitive units for each of the tested variables by side. We tested for independence between side and proportion of significant units by side using a chi-squared independence test (1 degree of freedom, p-values indicated for each variable), and found no evidence of lateralization for any of the tested variables.

	BL (n=117)	CM (n=39)	R (n=46)	p-value
Trial type	54	20	26	0.476
EV (exp)	12	2	5	0.590
EV (obs)	15	1	2	0.070
Outcome (exp)	12	2	8	0.184
Outcome (obs)	8	0	2	0.228
RPE (exp)	7	1	5	0.284
RPE (obs)	4	0	4	0.110

Table 4.2: Sensitive units by major subnuclei group. Number of sensitive units for each of the tested variables by major amygdalar subnuclei group (BL: deep and basolateral; CM: superficial or corticomedial; R: remaining nuclei). We tested for independence between major amygdalar subnuclei groups and proportion of significant units with a chi-squared test (2 degrees of freedom, p-values indicated for each variable), and found no evidence for any subnuclei group being more likely to contain sensitive units for any tested variable.

Decoding generalization analysis

To test whether the same or different neurons encode experienced and observational variables we performed a decoding generalization analysis (Wang, Mamelak, et al., 2019). We trained decoders with neural activity in experiential trials and tested in observational trials (Fig. 4.7a), and vice-versa (Fig. 4.7b). The method is otherwise

identical to the previous decoding analysis. We tested generalization of EVs (Fig. 4.7a,b, left panels) in the pre-outcome period and outcomes in the post-outcome period (Fig. 4.7a,b, right panels), since these variables were represented in the amygdala neuron population to some extent: outcome decoding was successful in both trial types, and despite weaker observational EV decoding, we did find a significant observational EV unit count through the encoding analysis.

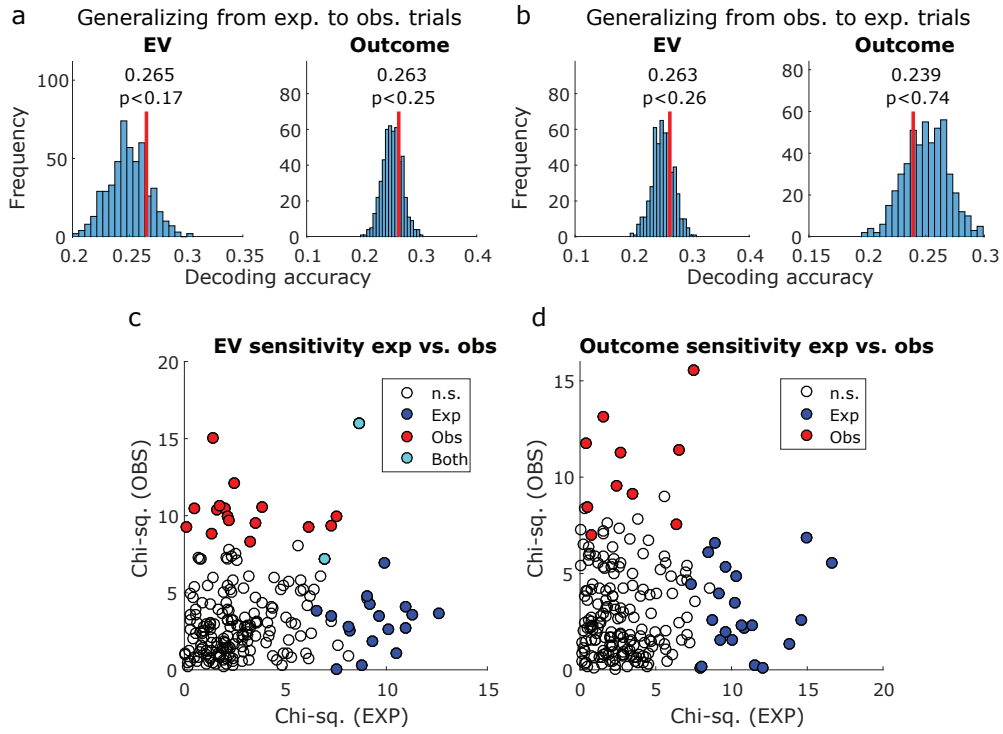

Figure 4.7: Comparing decoding and encoding across experiential and observational trials. (a) Decoding generalization, training a decoder in experiential trials and testing in observational trials. Decoded variables were EV (left) and outcome (right). Vertical red lines indicate decoding accuracy, and histograms indicate decoding accuracy in each instance of the permutation test with shuffled variable labels. P-values were obtained by computing the proportion of permutation iterations in which the decoding accuracy exceeded the true decoding accuracy. (b) Same, but training in observational trials and testing in experiential trials. (c) Sensitivity to EV in each unit, as obtained in the encoding analysis, plotted for experiential trials (x axis) and observational trials (y axis). Sensitivity was defined as the chi-squared value obtained from the Kruskal-Wallis test used in the encoding analysis. Unfilled data points indicate not sensitive units, blue data points indicate units only sensitive to experiential EV, red data points indicate units only sensitive to observational EV, and cyan data points indicate units sensitive to both experiential and observational EV. (b) Same, but for outcome sensitivity.

None of the generalization decoding tests yielded better than chance decoding

accuracy in the permutation test, regardless of which set of trials (experiential or observational) was used to train or test the decoder.

Additionally, we plotted the sensitivity of each individual amygdala neuron to EVs (pre-outcome, Fig. 4.7c) and outcomes (post-outcome, Fig. 4.7d), contrasting experiential and observational trials. The sensitivity of each neuron is defined as the chi-squared value obtained from the previously described encoding Kruskal-Wallis test for differing levels of EV or outcome. The Pearson correlation between EV sensitivities was $\rho = 0.10$ ($p < 0.14$), and only 2 units were found to be sensitive in both trial types. Additionally, the Pearson correlation between outcome sensitivities was $\rho = 0.02$ ($p < 0.77$), and no units were found to be sensitive in both trial types.

These results indicate that despite evidence for successful outcome decoding in each condition separately, and significant unit counts for EV in both conditions, there is no evidence supporting a shared representation between experiential and observational trial conditions in amygdala.

Simulation for comparing encoding and decoding

We note that for some of the variables we investigated, the results of the encoding and decoding analysis differ such that one was above chance where the other was not (see discussion for an interpretation of this finding). To better understand how encoding and decoding analyses might differ in our data, we simulated the performance of these methods in characterizing the activity of an artificial neuron population whose activity correlates with an artificial 4-category variable, with varying levels of noise. We used classification accuracy in test trials as the decoding metric, as well as the ratio of significant units as the encoding metric. For each noise factor level in $[1, 5, 20]$, we ran 100 independent simulations, and plotted decoding versus encoding performance for each simulation (Fig. 4.5h), as well as the theoretically estimated chance levels of 0.295 for decoding accuracy, and 0.075 for significant unit ratios.

For the lowest noise factor, 1, decoding and encoding performances were highest, and always above the estimated chance levels for both analyses. The mean decoding accuracy was 0.52 ± 0.006, and the mean ratio of significant units was 0.50 ± 0.004. For a noise factor of 5, however, only 52% of the simulations performed better than the chance level for decoding, even though all simulations performed better than chance in the encoding analysis. The mean decoding accuracy was 0.29 ± 0.005, and the mean ratio of significant units was 0.24 ± 0.003. Finally, for the highest

noise factor, 20, in 3% of simulations both decoding and encoding performed better than chance; in 19% of simulations decoding performed better than chance and encoding performed below chance levels; in 12% of simulations decoding was below chance level and encoding performed better than chance; and in 66% of simulations both analyses performed below chance. The mean decoding accuracy was 0.25 ± 0.004, and the mean ratio of significant units was 0.061 ± 0.001. This result, particularly in the high noise condition, suggests that there is a diversity of neural population configurations in which decoding analyses might detect the presence of information for a variable of interest to an acceptable degree, but single unit counts might be below chance thresholds. Conversely, it is also possible that an above-threshold count of significant units might not translate into successful population decoding with the chosen method. Overall, this indicates that these analyses are complementary when trying to understand the information contained in single neuron data, particularly in noisier conditions.

4.5 Discussion

In observational learning, an individual learns about the value of stimuli in the world not through direct experience, but instead through observing the experiences of others. Here we investigated whether the human amygdala contains neuronal representations of key computational variables relevant for learning about the value of stimuli through observation. We found evidence for the encoding of the EV of a stimulus in amygdala neurons, at the time when participants are observing another agent choose that stimulus before this agent received an outcome, even though on those specific trials no tangible reward outcome is obtained by the participant themselves. In addition, we found evidence that the amygdala contains decodable representations of outcomes during observational learning and experiential learning. Together, these results suggest that the human amygdala tracks several key reinforcement learning variables that can be deployed for observational reward-learning.

In addition, human amygdala neurons also strongly discriminated between whether or not a particular trial involved observational or experiential learning at the trial onset. This was the most robust signal found in the amygdala neurons, though this could at least in part be an effect of the distinct visual properties of experiential and observational trials (i.e. the presence of the face of the observed person, which can modulate amygdala cells (Minxha, C. Mosher, et al., 2017)), or of the blocked task design. Still, taken together with the RL computations found in the amygdala that were related to observational learning, these findings support a contribution of the

human amygdala to observational learning.

Consistent with a large literature describing the role of amygdala in anticipating rewards (Belova, Paton, and Salzman, 2008; Prévost, McNamee, et al., 2013; O'Doherty, Cockburn, and Pauli, 2017), we found evidence for experiential EV in both the single unit encoding analysis and the population decoding analysis, as well as for observational EV, in the single unit encoding analysis, further supporting the computational model as a meaningful description of behavior. Our findings in the experiential condition are compatible with results reported in monkey amygdala for expected values in the context of anticipating rewards for exploratory decision making (Costa, Mitz, and B. B. Averbeck, 2019).

An issue that requires further investigation is whether neurons encode experiential and observational expected value signals independently of the identity of the presented stimulus. Previous studies have reported stimulus identity encoding at the single neuron level in amygdala, such as the identity of faces (Gothard et al., 2007), visual categories (Fried, MacDonald, and C. L. Wilson, 1997; Kreiman, Koch, and Fried, 2000; Rutishauser, Ye, et al., 2015), but also identity-independent stimulus feature encoding, such as in familiarity/novelty tuning during memory retrieval (Rutishauser, Ye, et al., 2015) and ambiguity tuning during decision making (Wang, R. Yu, et al., 2017).

A related question is whether the neural substrate representing value in amygdala neurons is the same or different for observational and experiential learning. That is, do EVs and outcomes activate amygdala neurons in a similar manner, whether it occurs in an observational learning situation or an experiential learning situation? To test this, we trained a classifier to decode these variables in observational learning and tested this classifier on the same neurons during the experiential learning condition and vice-versa. In both cases we could not successfully decode signals when training on one condition and testing on the other. These findings suggest that neuronal coding of observational learning EV and outcomes is distinct and not-overlapping with the neuronal code for experiential learning prediction errors. Additionally, we inspected the sensitivity of individual neurons while encoding EVs and outcomes, and found that across the amygdala neuron population, experiential and observational sensitivities to these variables do not correlate. There is also little overlap between which neurons encode EV and outcomes in each condition. This does not preclude the existence of a distinct population of amygdala neurons, not found by this study, which encodes both experiential and observational values, as re-

ported elsewhere (S. W. Chang et al., 2015). Our findings also support the argument that the subjects properly understood the task and knew that the observed rewards would not be given to them. If this were not the case, the neural representation for expected values and outcomes likely would not be separate.

The encoding and decoding analyses gave slightly divergent results for some of the variables in the observational learning condition. For instance, expected value signals during observational learning were detected at levels higher than chance in the single-unit encoding analysis but not in the decoding analysis (with the decoding accuracy bordering but not reaching statistical significance). Such divergent results can arise due to differences in the nature of the neural signals being detected by the two methods. Encoding analysis assesses information encoded on average by individual neurons, whereas decoding analysis assesses whether information can be read out at the population level in individual trials. It is possible to decode from a population from which individual neurons cannot be selected at levels above those expected by chance (as we reported for outcome in observational trials) if the underlying code is distributed (Rumelhart, McClelland, and PDP Research Group, 1986; Rogers and McClelland, 2014) and/or exhibits correlated variability between neurons (Stefanini et al., 2019). This is because from the point of view of a decoder, neurons that by themselves are not informative can still be useful in the context of the population. This has been demonstrated experimentally: a study on the distributed encoding of space in rodents (Stefanini et al., 2019) showed that cells which individually do not provide a significant amount of information were nevertheless highly informative at the population level as demonstrated by a high importance index. Conversely, it is possible that neurons can be selected at proportions higher than expected by chance while not being able to decode from individual trials. This may happen in a scenario where several units are considered sensitive but weakly so, not providing enough information for single-trial decoding. By simulating encoding and decoding in an artificial population of neurons, we showed situations where either of these discrepancies between encoding and decoding analyses are possible, depending on the levels of noise used in the simulation.

In the present study we did not assess whether the observational learning signals we found in the amygdala are specifically recruited when observing another human agent, or rather are recruited when observing causal relationships between stimuli, actions, and outcomes irrespective of the nature of the agent performing the actions.

Thus, our design does not control for the social vs. non-social learning component of observational learning. An important direction for future studies would be to compare and contrast neuronal effects in the amygdala during observational learning when the agent is human or a computer. There is no strong reason to assume a-priori that the responses detected in the amygdala should be specific only to observed human agents. However, it is possible that the presence of a human might enhance the salience of the observed stimuli compared to the situation where the agent is non-human, which could potentially increase the magnitude of neuronal responses. Adding to this argument, a study using a modified dictator game in monkeys found that amygdala neurons mirrored value representations between rewards received by oneself and given to others, but no such mirroring was observed when a computer was responsible for delivering rewards to another monkey (S. W. Chang et al., 2015).

One important caveat is that proportions of sensitive amygdala neurons for value related variables have been higher in the monkey literature (S. W. Chang et al., 2015; Costa, Mitz, and B. B. Averbeck, 2019) than in the present study. However, there are many differences between species, recording techniques and task preparation that could lead to such differences in encoding proportions. Unlike in animal studies, our participants performed the task for less than one hour with no training, whereas training in animals is typically weeks to even months. Additionally, our recording electrodes were chronically implanted and could not be moved to search for responsive neurons, thereby providing an unbiased estimate. Finally, it is plausible that neuronal representations in the human amygdala are more complex, processing rich forms of information such as social networks or deep and elaborate knowledge about stimuli in the world and their associated values, meaning that seeing a relatively smaller proportion of neurons dedicated to value coding would not be entirely surprising.

To conclude, our findings support a role for the human amygdala in observational learning, particularly under situations where associations between stimuli and outcomes are learned about through observing the experiences of another agent. The amygdala was found to contain neuronal representations depicting the expected future reward associated with particular stimuli when observing the experiences of another agent interacting with and obtaining rewards from those stimuli. Furthermore, amygdala neuron populations contained decodable information for outcomes whether the subject experienced them or passively observed another agent receiving them. The specific contributions we have uncovered for the amygdala in obser-

vational learning adds to a burgeoning literature highlighting a broad role for this structure in social cognition more generally. (Adolphs, Tranel, and Damasio, 1998; Gothard et al., 2007; Adolphs, 2010; S. W. Chang et al., 2015; Minxha, C. Mosher, et al., 2017; Taubert et al., 2018).

4.6 Author contributions

T.G.A. performed spike data pre-processing and all subsequent data analyses, and wrote the manuscript; J.M. collected data, performed spike data pre-processing and wrote the manuscript; S.D. created the original task design; I.B.R. performed patient surgery; A.N.M. performed patient surgery; U.R. and J.O.D. both conceived and directed the project, and wrote the manuscript.

Chapter 5

CONCLUSION

5.1 Summary of results

In this thesis, I presented several new results on how the human brain performs value-based learning and decision-making, leveraging rare single neuron recordings from epilepsy patients to probe the neural implementation of these processes at a previously unattainable level of spatial and temporal resolution. Crucially, our approach was to combine electrophysiological measures with computational models of learning, with a specific focus on reinforcement learning models, to provide a theoretical foundation for our behavioral and neural hypotheses. This approach allowed us to test several hypotheses about how the key variables underlying value-based learning are represented and manipulated in neural populations — that is, we were able to probe the algorithmic level implementation of these processes with regards to Marr's levels of computational descriptions (Marr, 1982).

Our first set of results concern the ability of neural populations to integrate distinct value-based features for different decision options, producing an integrated utility value for stimuli. Such utility values for different stimuli could, in turn, be compared to produce a decision output with the goal of maximizing rewards under uncertainty. While feature integration for value construction has been investigated by previous neural studies, for components of nutritional value (Suzuki, Cross, and O'Doherty, 2017), and subjective artistic value (Iigaya et al., 2020), our study focused on key features for exploratory decision making: expected values, novelty and uncertainty, using a two-armed bandit gambling task. Our computational analysis of behavior revealed that these variables had separable contributions over patients' decisions: while patients learned to seek out options which had been rewarding in the past and thus chose higher expected value options more often, novelty and uncertainty produced opposite biases in option selection. Specifically, patients tended to choose a stimulus more often if it was the higher novelty option and less often if it was the higher uncertainty option. Additionally, we found that the behavioral model that best explained patients' decisions was one in which uncertainty and novelty interacted non-linearly prior to being integrated into stimulus utilities along with expected values, mirroring behavioral results found in a neurotypical population (Cockburn

et al., 2021). Neurally, we mapped stimulus utility integration and decision making to preSMA: following stimulus presentation, preSMA neural activity correlated with the individual components of stimulus value (expected values, uncertainty and novelty) and with the final decision output. We also found that a significant portion of value coding preSMA neurons had their activity better explained by an integrated utility model, suggesting that neural activity in this brain area can be a substrate for value-based decision making. Additionally, a dimensionality reduction analysis revealed that the preSMA neural population performed distributed encoding of stimulus utilities, followed by decisions. In the vmPFC, on the other hand, we found that all the value-based features, as well as the integrated utility signal, were predominantly encoded for the selected stimulus. This suggests a disparity between this region and preSMA during exploratory decision making; while preSMA activity encodes stimulus-specific features which could be used to drive value-based exploratory decisions, vmPFC reports on the consequences of the decision that was made.

Our second study investigated the neural implementation of model-based Pavlovian conditioning by human neural populations. Model-based learning, as opposed to model-free learning, is a class of learning algorithms in which goal-directed planning occurs, often based on explicit knowledge of a cognitive map between the possible states of the environment. While the neural implementation of this mode of learning has been documented in instrumental conditioning (O'Doherty, Cockburn, and Pauli, 2017), how it occurs in the absence of decision input, during Pavlovian conditioning, is less understood (Dayan and Berridge, 2014). Following neuroimaging results which linked parts of prefrontal cortex and amygdala to model-based value coding during Pavlovian learning (Prévost, McNamee, et al., 2013; Pauli, Gentile, et al., 2019), we used a two-step Pavlovian learning task to probe whether model-based representations of stimulus value and identity could be found in single neurons across these brain areas. Specifically, we hypothesized that model-based Pavlovian conditioning would entail stimulus-stimulus associations and predictive identity coding. In other words, if the algorithmic implementation of learning relies on a learning a complete cognitive map of environmental states, then at the time a stimulus is seen, neurons might already encode the identity of future predicted stimuli. Indeed, while both amygdala and vmPFC neurons performed expected value coding, we found evidence in vmPFC neural populations for predictive identity coding as predicted by a model-based algorithm of Pavlovian conditioning. Interestingly, the temporal precedence of spikes between simultaneously recorded

vmPFC and amygdala neuron pairs changed as a function of the expected value of a stimulus: when exposed to stimuli that were expected to lead to rewards, vmPFC spikes tended to precede amygdala spikes, and the opposite was true for stimuli that were expected to lead to no rewards. Taken together, these results shed light on how the human brain leverages cognitive maps to perform the algorithmic implementation of Pavlovian conditioning, above and beyond the application of simple stimulus-reward associations alone.

Finally, our third study expands on Pavlovian conditioning, focusing on the intersection between value-based learning and observational learning. Specifically, we used a bandit task with two alternatives to investigate how value-based learning occurred in amygdala neurons when patients acquired information from playing the game as opposed to when they learned exclusively from watching another agent play the game. A model comparison approach revealed that, among the hypothesized mechanisms, the most likely explanation for patients' behavior in the task was that they utilized a counterfactual learning mechanism, in which learning about one option simultaneously updated the other option in the opposite direction as well, reflecting the true structure of the task. While this task was not designed to investigate the creation of cognitive maps or the comparison between model-free and model-based learning modes, the counterfactual heuristic does suggest an algorithm which violates what would be possible with a pure model-free learning algorithm. Neurally, we found that amygdala neurons encoded expected rewards for oneself and others, but that mostly distinct amygdala populations performed these two functions. Additionally, we successfully decoded reward values from amygdala populations, but obtained clearly distinct neural activity patterns for experiential trials as opposed to observational trials. These findings provide new insight on the neural implementation of value-based learning, at the intersection with social cognition.

5.2 The relevance of joint behavioral and electrophysiological studies

Broadly speaking, the results of three studies discussed in this thesis can be jointly put forth as an argument for the value of considering behavior front and center along with neural data in the effort to understand human cognition. Indeed, the understanding afforded by rare opportunities such as recording invasively from refractory epilepsy is significantly amplified in light of the variables and theoretical constructs contained in computational models of behavior, which provide a valuable guide for mapping cognition at the intersection between Marr's computational and algorithmic levels of

description (Marr, 1982). In other words, a deeper understanding can be extracted from neural data when our efforts are accompanied by a theory of what function a neural system aims to accomplish, and by constructing and validating models for the algorithmic implementation of these processes. Specifically, only with a theory of how behavior is affected by utility integration or information-based variables were we able to make novel conclusions about the nature of value coding and decision making in preSMA and vmPFC neurons. Similarly, our theoretically grounded assumptions for model-based learning during Pavlovian conditioning eventually led us to specific neural hypotheses about how predictive identity coding must occur, supported by behavioral measures of pupil diameter and stimulus preference. There are several other examples of the deeper understanding afforded to neuroscientific analysis by behavioral measures (Niv, 2021), and it has been argued that one of the functions of neuroscience in the broader cognitive sciences is to provide constraints for cognitive theories of representation and computation (Cushman, 2020). While an artificial agent may employ many distinct efficient strategies to solve a given learning problem, the human brain presumably only engages with a subset of these strategies at a time. Indeed, even finding a computational strategy which works extremely well to solve a given learning problem *in silico* is insufficient to understand human cognition, as artificial agents start to achieve superhuman performances in a variety of learning and decision making tasks (Mnih et al., 2013; Silver et al., 2016). In this sense, neural evidence can provide a guide to arbitrate between different candidate theories for how cognitive strategies are implemented, taking into account the evolutionary and developmental constraints which led to the human brain as it is today.

Why does it matter to know the exact implementations through which the brain solves value learning problems to perform decisions? While I would argue this question is by itself a worthwhile pursuit from the standpoint of basic science research, there are important potential applications for understanding the cognition behind value learning, especially where brain health is concerned, which I will briefly discuss. Broadly speaking, the nascent field of computational psychiatry is engaged in applying the computational descriptions of how the brain represents and processes information to assist in critical endeavors for the field, such as finding biomarkers for mental illness and potential treatment routes (Redish and Gordon, 2016). For context, efforts have been made in creating computational models for addiction, schizophrenia, depression, obsessive-compulsive disorder (OCD), and autism, among other mental illnesses (Maia and Frank, 2011; Montague et al., 2012). For

instance, one computational framework proposes that addiction may partially result from maladaptive patterns involving points of failure (e.g. overvaluation of the habitual decision making system, or ineffective arbitration between model-free and model-based decisions) in the same corticostriatal circuitry implicated in reinforcement learning (Redish, Jensen, and Johnson, 2008). In schizophrenia, which may include symptoms ranging from hallucinations to anhedonia, a commonly reported neurological trait is a dopaminergic imbalance, expressed by excessive dopamine in striatum and reduced dopamine in PFC (Guillin, Abi-Dargham, and Laruelle, 2007). Given that one of the proposed computational roles of dopamine is to signal the salience of rewards (Berridge and T. E. Robinson, 1998), a hypothetical link was suggested between the observed psychosis and the heightened underlying salience, or *aberrant salience* of stimuli, caused by dysfunctional dopaminergic expression (Kapur, 2003; Howes et al., 2020). Additionally, converging evidence has implicated a corticostriatal circuitry involving OFC and ACC in OCD (Maia, Cooney, and Peterson, 2008), given that these regions tend to be hyperactive in patients relative to controls, and that this disparity is reduced following treatment. Given the role of these regions in model-based control, some studies hypothesized that OCD patients would exhibit disparities relative to controls in arbitrating between model-free and model-based control (Voon, Reiter, et al., 2017). Indeed, the degree of compulsivity in patients was found to correlate negatively with model-based control and positively with model-free control (Voon, Baek, et al., 2015). In studies of mood disorders, for instance, a computational framework has been proposed to explain *learned helplessness*, a key feature of depression: it can be viewed as a probabilistic account of environments in which an individual's decisions have no predictive power over the punishments and rewards it receives (Montague et al., 2012). This is congruent with how this type of response can be experimentally induced in animals, through unpredictable reward regimes (Goodkin, 1976).

5.3 Discussion and future research directions

Our results suggest a number of possible future research directions in human electrophysiology and value-based decision making. As I previously discussed, dlPFC is a brain area which provides a cortical bridge between the value-coding prefrontal regions and preSMA (Luppino et al., 1993), which we found to encode an integrated utility signal using value-based variables such as uncertainty and novelty as components, as well as decisions themselves. Together with our finding that vmPFC predominantly encoded integrated utilities conditioned on decisions, unlike

preSMA, which encoded pre-decision components of utility for individual stimuli, this raises a number of questions on the functional organization of prefrontal cortex. Previous neuroimaging evidence did implicate vmPFC/medial OFC in encoding integrated utility values and suggested this integration can leverage components encoded in areas such as lateral OFC and posterior parietal cortex (Suzuki, Cross, and O'Doherty, 2017; Iigaya et al., 2020). Is it possible that this discrepancy arises from a difference in the nature of the tasks utilized? In the aforementioned examples, subjects performed behavioral paradigms mostly focused on valuation alone (e.g. willingness-to-pay for food; subjective art appraisals), while our explore-exploit paradigm encouraged subjects to accumulate value-based evidence to perform reward maximizing decisions under uncertainty. Therefore, one hypothesis is that exploratory decision making induces pre-decision value integration to occur more prominently in either dlPFC or preSMA than in vmPFC/OFC, which then receive utility feedback from the more posterior value regions. Also note that dlPFC connects recurrently to preSMA, vmPFC, and OFC in such a way that a one-directional cascade of events across these brain regions, while appealing, is likely a functional oversimplification of this system. Another possibility is that these results are actually congruent, and the previously described role of vmPFC in encoding integrated utility concerns selected stimuli rather than individual options being considered prior to a decision, as these paradigms did not have a multi-option selection component to disambiguate the two possibilities. One key testable hypothesis is that utility integration prior to value-based decision first occurs in dlPFC rather than in preSMA, which we could not probe due to limitations in choosing recording sites in human patients. Admittedly, it is also possible that pre-decision utility integration during exploration occurs in a different subset of vmPFC neurons, which we simply did not sample.

While the present thesis focused predominantly on a number of cortical areas and cortico-cortical relationships, it is crucial to note that value learning and decision making take place in a much broader circuitry involving subcortical areas such as substantia nigra, ventral tegmental area, subthalamic nucleus, striatum and thalamus (O'Doherty, Cockburn, and Pauli, 2017). For instance, the striatum receives topographically organized cortical inputs from motor and premotor areas, OFC, and ACC, and is also one of the main targets for dopaminergic neurons from SN/VTA, which are known to encode reward prediction errors (Haber and Knutson, 2010). Activity in human ventral striatum reflected model-free encoding of expected rewards (Tobler et al., 2006), while dorsomedial striatum in rodents is necessary

for goal-directed learning to take place (Yin et al., 2005) altogether. Additionally, evidence from monkey electrophysiology suggests that spatial representations in striatal activity are least partially modulated by uncertainty (Yanike and Ferrera, 2014). The striatum receives projections from thalamus but also indirectly projects back to it, mediated by either the internal segment of the globus pallidus (GPi), in the *direct pathway*, or the external segment of the globus pallidus (GPe), followed by the subthalamic nucleus (STN) and then GPi, in the *indirect pathway* (Haber, 2016). The thalamus then projects back to cortex, thus completing the cortical-striatal-thalamic loop (Haber and Knutson, 2010). Specifically, the medial-dorsal nucleus projects substantially to OFC and dlPFC, while the ventral-anterior nucleus projects to preSMA (McFarland and Haber, 2002). The STN also receives topographically organized projections of special interest from value coding areas of cortex (including vmPFC/OFC, dlPFC, and dACC), which define the *hyperdirect pathway* (Haynes and Haber, 2013). Mapping the role of the human hyperdirect pathway in value learning and decision making will be a crucial endeavor for the field. Recent deep-brain stimulation work in Parkinson's disease patients has allowed investigators to probe these subcortical targets in the human brain directly (Pouratian et al., 2012; C. P. Mosher et al., 2021), creating an exciting new venue for reward learning studies to take place.

Additionally, while our findings support a model-based account for Pavlovian conditioning and the usage of learning heuristics that go beyond model-free learning in the context of social cognition, the present work did not sufficiently explore the neural implementation of arbitration between multiple controllers, such as model-free vs. model-based learning algorithms for value learning or emulation vs. imitation strategies for observational learning. On the one hand, we were able to measure evidence for both reward prediction error and state prediction error signals in preSMA, which can serve as a substrate for computing reliability signals to mediate the arbitration between different learning systems. On the other hand, we did not record from key areas suggested by neuroimaging studies to be directly involved in arbitrating multiple controllers, such as inferior lateral PFC (S. W. Lee, Shimojo, and O'Doherty, 2014), which could be a potential future target for local field potential (LFP) recordings.

Finally, the field of reinforcement learning itself has rapidly developed in the recent years, providing a theoretical basis to explain more complex phenomena, opening new possibilities for neuroscience to explore. For instance, distributional RL

proposes a framework for learning entire distributions for probabilistic variables, beyond expected values alone (Bellemare, Dabney, and Munos, 2017). This framework can be used to make neuroscientific predictions, some of which already have tentative evidence in their support. For example, one way to support distributional learning would be having a diversity of optimistic and pessimistic dopaminergic neurons, which collectively generate a distribution of reward prediction error for the same outcome; this prediction has found initial support from VTA recordings in rodents (Dabney, Kurth-Nelson, et al., 2020). Other implications from this theory still must be tested, especially as far as the human brain is concerned, including how the brain leverages the knowledge entire reward distributions to make decisions, beyond their expected value, or how distributional learning interfaces with model-based reward learning and goal-directed planning (Lowet et al., 2020). Additionally, deep reinforcement learning has developed as a method to provide new solutions to complex problem solving in high dimensional environments (Mnih et al., 2013; Silver et al., 2016). While deep neural networks have been successfully used as a structural and functional model for the processing of visual information (Yamins and DiCarlo, 2016), whether deep reinforcement learning models are efficient at predicting decision-making activity in the human brain is an active research area, in which encouraging results started to be established, connecting artificial value representations in neural activity across cortex (Cross et al., 2021).

BIBLIOGRAPHY

Adolphs, Ralph (2010). "What does the amygdala contribute to social cognition?" In: *Annals of the New York Academy of Sciences* 1191.1, pp. 42–61.

Adolphs, Ralph, Daniel Tranel, and Antonio R Damasio (1998). "The human amygdala in social judgment". In: *Nature* 393.6684, p. 470.

Aggleton, JP, MJ Burton, and RE Passingham (1980). "Cortical and subcortical afferents to the amygdala of the rhesus monkey (Macaca mulatta)". In: *Brain Research* 190.2, pp. 347–368.

Allsop, Stephen A et al. (2018). "Corticoamygdala transfer of socially derived information gates observational learning". In: *Cell* 173.6, pp. 1329–1342.

Amaral, David G (1986). "Amygdalohippocampal and amygdalocortical projections in the primate brain". In: *Excitatory amino acids and epilepsy*. Springer, pp. 3–17.

Amiez, Céline, Jean-Paul Joseph, and Emmanuel Procyk (2006). "Reward encoding in the monkey anterior cingulate cortex". In: *Cerebral Cortex* 16.7, pp. 1040–1055.

Applegate, Craig D et al. (1982). "Multiple unit activity recorded from amygdala central nucleus during Pavlovian heart rate conditioning in rabbit". In: *Brain research* 238.2, pp. 457–462.

Aquino, Tomas G. et al. (2021). "Neurons in human pre-supplementary motor area encode key computations for value-based choice". In: *bioRxiv*. DOI: 10.1101/2021.10.27.466000.

Arcediano, Francisco, Helena Matute, and Ralph R Miller (1997). "Blocking of Pavlovian conditioning in humans". In: *Learning and Motivation* 28.2, pp. 188–199.

Averbeck, Bruno and John P O'Doherty (2021). "Reinforcement-learning in fronto-striatal circuits". In: *Neuropsychopharmacology*, pp. 1–16.

Badre, David et al. (2012). "Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration". In: *Neuron* 73.3, pp. 595–607.

Balleine, Bernard W and Anthony Dickinson (1998). "Goal-directed instrumental action: contingency and incentive learning and their cortical substrates". In: *Neuropharmacology* 37.4-5, pp. 407–419.

Barraclough, Dominic J, Michelle L Conroy, and Daeyeol Lee (2004). "Prefrontal cortex and decision making in a mixed-strategy game". In: *Nature neuroscience* 7.4, pp. 404–410.

Baxter, Mark G et al. (2000). "Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex". In: *Journal of Neuroscience* 20.11, pp. 4311–4319.

Bechara, Antoine et al. (1999). "Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making". In: *Journal of Neuroscience* 19.13, pp. 5473–5481.

Bellemare, Marc G, Will Dabney, and Rémi Munos (2017). "A distributional perspective on reinforcement learning". In: *International Conference on Machine Learning*. PMLR, pp. 449–458.

Belova, Marina A, Joseph J Paton, Sara E Morrison, et al. (2007). "Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala". In: *Neuron* 55.6, pp. 970–984.

Belova, Marina A, Joseph J Paton, and C Daniel Salzman (2008). "Moment-to-moment tracking of state value in the amygdala". In: *Journal of Neuroscience* 28.40, pp. 10023–10030.

Berridge, Kent C and Terry E Robinson (1998). "What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?" In: *Brain research reviews* 28.3, pp. 309–369.

Bonini, Francesca et al. (2014). "Action monitoring and medial frontal cortex: leading role of supplementary motor area". In: *Science* 343.6173, pp. 888–891.

Bourgeois, Jean-Pierre et al. (2012). "Modulation of the mouse prefrontal cortex activation by neuronal nicotinic receptors during novelty exploration but not by exploration of a familiar environment". In: *Cerebral Cortex* 22.5, pp. 1007–1015.

Bozkurt, Ahmet et al. (2001). "Organization of primate amygdalo-prefrontal projections". In: *Neurocomputing* 38, pp. 1135–1140.

Brody, Carlos D (1999). "Correlations without synchrony". In: *Neural Computation* 11.7, pp. 1537–1551.

Brogden, Wilfred J (1947). "Sensory preconditioning of human subjects." In: *Journal of experimental psychology* 37.6, p. 527.

Brown, Paul L and Herbert M Jenkins (1968). "Auto-shaping of the pigeon's key peck". In: *Journal of the experimental analysis of behavior* 11.1, pp. 1–8.

Cai, Xinying and Camillo Padoa-Schioppa (2012). "Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex". In: *Journal of Neuroscience* 32.11, pp. 3791–3808.

Carcea, Ioana and Robert C Froemke (2019). "Biological mechanisms for observational learning". In: *Current opinion in neurobiology* 54, pp. 178–185.

Carmichael, S Thomas and JL Price (1996). "Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys". In: *Journal of Comparative Neurology* 371.2, pp. 179–207.

Chang, Steve WC et al. (2015). "Neural mechanisms of social decision-making in the primate amygdala". In: *Proceedings of the National Academy of Sciences* 112.52, pp. 16012–16017.

Charpentier, Caroline J, Kiyohito Iigaya, and John P O'Doherty (2020). "A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning". In: *Neuron* 106.4, pp. 687–699.

Charpentier, Caroline J and John P O'Doherty (2018). "The application of computational models to social neuroscience: promises and pitfalls". In: *Social Neuroscience* 13.6, pp. 637–647.

Chen, Witney et al. (2020). "Prefrontal-subthalamic hyperdirect pathway modulates movement inhibition in humans". In: *Neuron* 106.4, pp. 579–588.

Chib, Vikram S et al. (2009). "Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex". In: *Journal of Neuroscience* 29.39, pp. 12315–12320.

Cockburn, Jeffrey et al. (2021). "Novelty and uncertainty interact to regulate the balance between exploration and exploitation in the human brain." In: *bioRxiv*. DOI: 10.1101/2021.10.13.464279. eprint: https://www.biorxiv.org/content/early/2021/10/14/2021.10.13.464279.full.pdf. URL: https://www.biorxiv.org/content/early/2021/10/14/2021.10.13.464279.

Cohen, Jonathan D, Samuel M McClure, and Angela J Yu (2007). "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481, pp. 933–942.

Cooper, Jeffrey C et al. (2012). "Human dorsal striatum encodes prediction errors during observational learning of instrumental actions". In: *Journal of Cognitive Neuroscience* 24.1, pp. 106–118.

Corbit, Laura H and Bernard W Balleine (2005). "Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer". In: *Journal of Neuroscience* 25.4, pp. 962–970.

Costa, Vincent D, Andrew R Mitz, and Bruno B Averbeck (2019). "Subcortical substrates of explore-exploit decisions in primates". In: *Neuron* 103.3, pp. 533–545.

Cross, Logan et al. (2021). "Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments". In: *Neuron* 109.4, pp. 724–738.

Cushman, Fiery (2020). "Is cognitive neuroscience an oxymoron". In: *Current controversies in philosophy of cognitive science*, pp. 121–133.

D'Ardenne, Kimberlee et al. (2008). "BOLD responses reflecting dopaminergic signals in the human ventral tegmental area". In: *Science* 319.5867, pp. 1264–1267.

Dabney, Will, Zeb Kurth-Nelson, et al. (2020). "A distributional code for value in dopamine-based reinforcement learning". In: *Nature* 577.7792, pp. 671–675.

Dabney, Will, Mark Rowland, et al. (2018). "Distributional reinforcement learning with quantile regression". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Davey, Graham CL (1992). "Classical conditioning and the acquisition of human fears and phobias: A review and synthesis of the literature". In: *Advances in Behaviour Research and Therapy* 14.1, pp. 29–66.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature Neuroscience* 8.12, p. 1704.

Dayan, Peter and Kent C Berridge (2014). "Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation". In: *Cognitive, Affective, & Behavioral Neuroscience* 14.2, pp. 473–492.

De Araujo, Ivan ET, Morten L Kringelbach, et al. (2003). "Human cortical responses to water in the mouth, and the effects of thirst". In: *Journal of neurophysiology* 90.3, pp. 1865–1876.

De Araujo, Ivan ET, Edmund T Rolls, et al. (2003). "Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain". In: *European Journal of Neuroscience* 18.7, pp. 2059–2068.

De Martino, Benedetto, Colin F Camerer, and Ralph Adolphs (2010). "Amygdala damage eliminates monetary loss aversion". In: *Proceedings of the National Academy of Sciences* 107.8, pp. 3788–3792.

Dias, Rebecca and Robert Colin Honey (2002). "Involvement of the rat medial prefrontal cortex in novelty detection." In: *Behavioral Neuroscience* 116.3, p. 498.

Ding, Long and Joshua I Gold (2010). "Caudate encodes multiple computations for perceptual decisions". In: *Journal of Neuroscience* 30.47, pp. 15747–15759.

Domenech, Philippe and Etienne Koechlin (2015). "Executive control and decision-making in the prefrontal cortex". In: *Current Opinion in Behavioral Sciences* 1, pp. 101–106.

Domenech, Philippe, Sylvain Rheims, and Etienne Koechlin (2020). "Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex". In: *Science* 369.6507.

Doya, Kenji (1999). "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?" In: *Neural Networks* 12.7-8, pp. 961–974.

Du, Jingnan et al. (2020). "Functional connectivity of the orbitofrontal cortex, anterior cingulate cortex, and inferior frontal gyrus in humans". In: *Cortex* 123, pp. 185–199.

Dukas, Reuven and Elizabeth A Bernays (2000). "Learning improves growth rate in grasshoppers". In: *Proceedings of the National Academy of Sciences* 97.6, pp. 2637–2640.

Dunne, Simon, Arun D'Souza, and John P O'Doherty (2016). "The involvement of model-based but not model-free learning signals during observational reward learning in the absence of choice". In: *Journal of Neurophysiology* 115.6, pp. 3195–3203.

Fan, Yunshu, Joshua I Gold, and Long Ding (2020). "Frontal eye field and caudate neurons make different contributions to reward-biased perceptual decisions". In: *Elife* 9, e60535.

Fried, Itzhak, Katherine A MacDonald, and Charles L Wilson (1997). "Single neuron activity in human hippocampus and amygdala during recognition of faces and objects". In: *Neuron* 18.5, pp. 753–765.

Fried, Itzhak, Roy Mukamel, and Gabriel Kreiman (2011). "Internally generated preactivation of single neurons in human medial frontal cortex predicts volition". In: *Neuron* 69.3, pp. 548–562.

Fu, Zhongzheng et al. (2019). "Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal Cortex". In: *Neuron* 101.1, pp. 165–177.

Gazit, Tomer et al. (2020). "The role of mPFC and MTL neurons in human choice under goal-conflict". In: *Nature Communications* 11.1, pp. 1–12.

Gershman, Samuel J (2018). "Deconstructing the human algorithms for exploration". In: *Cognition* 173, pp. 34–42.

Ghashghaei, HT, Claus C Hilgetag, and Helen Barbas (2007). "Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala". In: *Neuroimage* 34.3, pp. 905–923.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Gnadt, James W and Richard A Andersen (1988). "Memory related motor planning activity in posterior parietal cortex of macaque". In: *Experimental Brain Research* 70.1, pp. 216–220.

Gold, Joshua I and Michael N Shadlen (2007). "The neural basis of decision making". In: *Annu. Rev. Neurosci.* 30, pp. 535–574.

Goñi, Joaquin et al. (2011). "The neural substrate and functional integration of uncertainty in decision making: an information theory approach". In: *PloS one* 6.3, e17408.

Goodkin, Franklin (1976). "Rats learn the relationship between responding and environmental events: An expansion of the learned helplessness hypothesis". In: *Learning and Motivation* 7.3, pp. 382–393.

Gothard, Katalin M et al. (2007). "Neural responses to facial expression and face identity in the monkey amygdala". In: *Journal of Neurophysiology*.

Gottfried, Jay A, John P O'Doherty, and Raymond J Dolan (2002). "Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging". In: *Journal of Neuroscience* 22.24, pp. 10829–10837.

– (2003). "Encoding predictive reward value in human amygdala and orbitofrontal cortex". In: *Science* 301.5636, pp. 1104–1107.

Grabenhorst, Fabian, Raymundo Báez-Mendoza, et al. (2019). "Primate Amygdala Neurons Simulate Decision Processes of Social Partners". In: *Cell*.

Grabenhorst, Fabian, István Hernádi, and Wolfram Schultz (2012). "Prediction of economic choice by primate amygdala neurons". In: *Proceedings of the National Academy of Sciences* 109.46, pp. 18950–18955.

Grabenhorst, Fabian and Edmund T Rolls (2011). "Value, pleasure and choice in the ventral prefrontal cortex". In: *Trends in Cognitive Sciences* 15.2, pp. 56–67.

Guillin, Olivier, Anissa Abi-Dargham, and Marc Laruelle (2007). "Neurobiology of dopamine in schizophrenia". In: *International review of neurobiology* 78, pp. 1–39.

Gutierrez, Ranier et al. (2006). "Orbitofrontal ensemble activity monitors licking and distinguishes among natural rewards". In: *Journal of Neurophysiology* 95.1, pp. 119–133.

Haber, Suzanne N (2016). "Corticostriatal circuitry". In: *Dialogues in clinical neuroscience* 18.1, p. 7.

Haber, Suzanne N and Brian Knutson (2010). "The reward circuit: linking primate anatomy and human imaging". In: *Neuropsychopharmacology* 35.1, pp. 4–26.

Hampton, Alan N, Ralph Adolphs, et al. (2007). "Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex". In: *Neuron* 55.4, pp. 545–555.

Hampton, Alan N, Peter Bossaerts, and John P O'Doherty (2006). "The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans". In: *Journal of Neuroscience* 26.32, pp. 8360–8367.

Hanes, Doug P, Kirk G Thompson, and Jeffrey D Schall (1995). "Relationship of presaccadic activity in frontal eye field and supplementary eye field to saccade initiation in macaque: Poisson spike train analysis". In: *Experimental Brain Research* 103.1, pp. 85–96.

Hare, Todd A, Shabnam Hakimi, and Antonio Rangel (2014). "Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting". In: *Frontiers in neuroscience* 8, p. 50.

Hare, Todd A, Wolfram Schultz, et al. (2011). "Transformation of stimulus value signals into motor commands during simple choice". In: *Proceedings of the National Academy of Sciences* 108.44, pp. 18120–18125.

Haynes, William IA and Suzanne N Haber (2013). "The organization of prefrontal-subthalamic inputs in primates provides an anatomical substrate for both functional specificity and integration: implications for Basal Ganglia models and deep brain stimulation". In: *Journal of Neuroscience* 33.11, pp. 4804–4814.

Heekeren, Hauke R et al. (2004). "A general mechanism for perceptual decision-making in the human brain". In: *Nature* 431.7010, pp. 859–862.

Hill, Michael R, Erie D Boorman, and Itzhak Fried (2016). "Observational learning computations in neurons of the human anterior cingulate cortex". In: *Nature Communications* 7, p. 12722.

Hirokawa, Junya et al. (2019). "Frontal cortex neuron types categorically encode single decision variables". In: *Nature* 576.7787, pp. 446–451.

Holland, Peter C and Michela Gallagher (2004). "Amygdala–frontal interactions and reward expectancy". In: *Current Opinion in Neurobiology* 14.2, pp. 148–155.

Holt, Gary R et al. (1996). "Comparison of discharge variability in vitro and in vivo in cat visual cortex neurons". In: *Journal of Neurophysiology* 75.5, pp. 1806–1814.

Horst, Nicole K and Mark Laubach (2013). "Reward-related activity in the medial prefrontal cortex is driven by consumption". In: *Frontiers in Neuroscience* 7, p. 56.

Horvitz, Jon C, Tripp Stewart, and Barry L Jacobs (1997). "Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat". In: *Brain Research* 759.2, pp. 251–258.

Howard, James D et al. (2015). "Identity-specific coding of future rewards in the human orbitofrontal cortex". In: *Proceedings of the National Academy of Sciences* 112.16, pp. 5195–5200.

Howes, Oliver D et al. (2020). "Aberrant salience, information processing, and dopaminergic signaling in people at clinical high risk for psychosis". In: *Biological psychiatry* 88.4, pp. 304–314.

Hunt, Laurence T et al. (2018). "Triple dissociation of attention and decision computations across prefrontal cortex". In: *Nature Neuroscience* 21.10, pp. 1471–1481.

Iacoboni, Marco et al. (1999). "Cortical mechanisms of human imitation". In: *science* 286.5449, pp. 2526–2528.

Iigaya, Kiyohito et al. (2020). "Aesthetic preference for art emerges from a weighted integration over hierarchically structured visual features in the brain". In: *BioRxiv*.

Jaccard, Paul (1912). "The distribution of the flora in the alpine zone. 1". In: *New Phytologist* 11.2, pp. 37–50.

Jenison, Rick L et al. (2011). “Value encoding in single neurons in the human amygdala during decision making”. In: *Journal of Neuroscience* 31.1, pp. 331–338.

Jones, Joshua L et al. (2012). “Orbitofrontal cortex supports behavior and learning using inferred but not cached values”. In: *Science* 338.6109, pp. 953–956.

Kahneman, Daniel (2011). *Thinking, fast and slow*. Macmillan.

Kahnt, Thorsten et al. (2010). “The neural code of reward anticipation in human orbitofrontal cortex”. In: *Proceedings of the National Academy of Sciences* 107.13, pp. 6010–6015.

– (2011). “Decoding different roles for vmPFC and dlPFC in multi-attribute decision making”. In: *Neuroimage* 56.2, pp. 709–715.

Kamiński, Jan et al. (2018). “Novelty-sensitive dopaminergic neurons in the human substantia nigra predict success of declarative memory formation”. In: *Current Biology* 28.9, pp. 1333–1343.

Kapur, Shitij (2003). “Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia”. In: *American journal of Psychiatry* 160.1, pp. 13–23.

Kennerley, Steven W, Timothy EJ Behrens, and Joni D Wallis (2011). “Double dissociation of value computations in orbitofrontal and anterior cingulate neurons”. In: *Nature Neuroscience* 14.12, pp. 1581–1589.

Kennerley, Steven W and Joni D Wallis (2009). “Encoding of reward and space during a working memory task in the orbitofrontal cortex and anterior cingulate sulcus”. In: *Journal of neurophysiology*.

Kepecs, Adam et al. (2008). “Neural correlates, computation and behavioural impact of decision confidence”. In: *Nature* 455.7210, pp. 227–231.

Kim, Jong-Nam and Michael N Shadlen (1999). “Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque”. In: *Nature Neuroscience* 2.2, pp. 176–185.

Klein-Flügge, Miriam Cornelia et al. (2013). “Segregated encoding of reward–identity and stimulus–reward associations in human orbitofrontal cortex”. In: *Journal of Neuroscience* 33.7, pp. 3202–3211.

Knudsen, Eric B and Joni D Wallis (2020). “Closed-loop theta stimulation in the orbitofrontal cortex prevents reward-based learning”. In: *Neuron* 106.3, pp. 537–547.

Kobak, Dmitry et al. (2016). “Demixed principal component analysis of neural population data”. In: *Elife* 5, e10989.

Kobayashi, Kenji and Ming Hsu (2019). “Common neural code for reward and information value”. In: *Proceedings of the National Academy of Sciences* 116.26, pp. 13061–13066.

Krajbich, Ian and Antonio Rangel (2011). "Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions". In: *Proceedings of the National Academy of Sciences* 108.33, pp. 13852–13857.

Krebs, Ruth M et al. (2009). "The novelty exploration bonus and its attentional modulation". In: *Neuropsychologia* 47.11, pp. 2272–2281.

Kreiman, Gabriel, Christof Koch, and Itzhak Fried (2000). "Category-specific visual responses of single neurons in the human medial temporal lobe". In: *Nature Neuroscience* 3.9, p. 946.

Kruskal, William H and W Allen Wallis (1952). "Use of ranks in one-criterion variance analysis". In: *Journal of the American statistical Association* 47.260, pp. 583–621.

Lee, Daeyeol, Hyojung Seo, and Min Whan Jung (2012). "Neural basis of reinforcement learning and decision making". In: *Annual Review of Neuroscience* 35, pp. 287–308.

Lee, Sang Wan, Shinsuke Shimojo, and John P O'Doherty (2014). "Neural computations underlying arbitration between model-based and model-free learning". In: *Neuron* 81.3, pp. 687–699.

Levy, Dino J and Paul W Glimcher (2012). "The root of all value: a neural common currency for choice". In: *Current Opinion in Neurobiology* 22.6, pp. 1027–1038.

Li, Yansong et al. (2016). "The neural dynamics of reward value and risk coding in the human orbitofrontal cortex". In: *Brain* 139.4, pp. 1295–1309.

Lichtenberg, Nina T et al. (2017). "Basolateral amygdala to orbitofrontal cortex projections enable cue-triggered reward expectations". In: *Journal of Neuroscience*, pp. 0486–17.

Lowet, Adam S et al. (2020). "Distributional reinforcement learning in the brain". In: *Trends in Neurosciences*.

Lubow, Robert E and Jacob C Gewirtz (1995). "Latent inhibition in humans: data, theory, and implications for schizophrenia." In: *Psychological bulletin* 117.1, p. 87.

Luppino, Giuseppe et al. (1993). "Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey". In: *Journal of Comparative Neurology* 338.1, pp. 114–140.

Maia, Tiago V, Rebecca E Cooney, and Bradley S Peterson (2008). "The neural bases of obsessive–compulsive disorder in children and adults". In: *Development and psychopathology* 20.4, pp. 1251–1283.

Maia, Tiago V and Michael J Frank (2011). "From reinforcement learning models to psychiatric and neurological disorders". In: *Nature neuroscience* 14.2, pp. 154–162.

Málková, Ludiše, David Gaffan, and Elisabeth A Murray (1997). "Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys". In: *Journal of Neuroscience* 17.15, pp. 6011–6020.

Malvaez, Melissa et al. (2019). "Distinct cortical–amygdala projections drive reward value encoding and retrieval". In: *Nature Neuroscience* 22.5, pp. 762–769.

Marr, David (1982). "Vision: A computational investigation into the human representation and processing of visual information, henry holt and co". In: *Inc., New York, NY* 2.4.2.

Matsumoto, Madoka, Kenji Matsumoto, Hiroshi Abe, et al. (2007). "Medial prefrontal cell activity signaling prediction errors of action values". In: *Nature Neuroscience* 10.5, pp. 647–656.

Matsumoto, Madoka, Kenji Matsumoto, and Keiji Tanaka (2007). "Effects of novelty on activity of lateral and medial prefrontal neurons". In: *Neuroscience research* 57.2, pp. 268–276.

McDannald, Michael A et al. (2011). "Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning". In: *Journal of Neuroscience* 31.7, pp. 2700–2705.

McFarland, Nikolaus R and Suzanne N Haber (2002). "Thalamic relay nuclei of the basal ganglia form both reciprocal and nonreciprocal cortical connections, linking multiple frontal cortical areas". In: *Journal of Neuroscience* 22.18, pp. 8117–8132.

McNamee, Daniel, Antonio Rangel, and John P O'Doherty (2013). "Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex". In: *Nature Neuroscience* 16.4, pp. 479–485.

Meyers, Ethan (2013). "The neural decoding toolbox". In: *Frontiers in Neuroinformatics* 7, p. 8.

Milosavljevic, Milica et al. (2010). "The drift diffusion model can account for value-based choice response times under high and low time pressure". In: *Judgment and Decision Making* 5.6, pp. 437–449.

Minxha, Juri, Ralph Adolphs, et al. (2020). "Flexible recruitment of memory-based choice representations by the human medial frontal cortex". In: *Science* 368.6498.

Minxha, Juri, Adam N Mamelak, and Ueli Rutishauser (2018). "Surgical and electrophysiological techniques for single-neuron recordings in human epilepsy patients". In: *Extracellular Recording Approaches*. Springer, pp. 267–293.

Minxha, Juri, Clayton Mosher, et al. (2017). "Fixations gate species-specific responses to free viewing of faces in the human and macaque amygdala". In: *Cell Reports* 18.4, pp. 878–891.

Mnih, Volodymyr et al. (2013). "Playing atari with deep reinforcement learning". In: *arXiv preprint arXiv:1312.5602*.

Montague, P Read et al. (2012). "Computational psychiatry". In: *Trends in cognitive sciences* 16.1, pp. 72–80.

Mooney, Richard (2014). "Auditory–vocal mirroring in songbirds". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1644, p. 20130179.

Morecraft, RJ, C Geula, and M-M Mesulam (1992). "Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey". In: *Journal of Comparative Neurology* 323.3, pp. 341–358.

Mosher, Clayton P et al. (2021). "Distinct roles of dorsal and ventral subthalamic neurons in action selection and cancellation". In: *Neuron* 109.5, pp. 869–881.

Murray, Elisabeth A (2007). "The amygdala, reward and emotion". In: *Trends in Cognitive Sciences* 11.11, pp. 489–497.

Murray, Elisabeth A and Peter H Rudebeck (2018). "Specializations for reward-guided decision-making in the primate ventral prefrontal cortex". In: *Nature Reviews Neuroscience* 19.7, pp. 404–417.

Nambu, Atsushi, Hironobu Tokuno, and Masahiko Takada (2002). "Functional significance of the cortico–subthalamo–pallidal 'hyperdirect'pathway". In: *Neuroscience Research* 43.2, pp. 111–117.

Niv, Yael (2021). "The primacy of behavioral research for understanding the brain." In: *Behavioral Neuroscience*.

Noonan, MP, RB Mars, and MFS Rushworth (2011). "Distinct roles of three frontal cortical areas in reward-guided behavior". In: *Journal of Neuroscience* 31.40, pp. 14399–14412.

O'Doherty, John P, Jeffrey Cockburn, and Wolfgang M Pauli (2017). "Learning, reward, and decision making". In: *Annual Review of Psychology* 68, pp. 73–100.

O'Doherty, John P, Peter Dayan, et al. (2003). "Temporal difference models and reward-related learning in the human brain". In: *Neuron* 38.2, pp. 329–337.

O'Neill, Martin and Wolfram Schultz (2010). "Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value". In: *Neuron* 68.4, pp. 789–800.

O'Doherty, John P et al. (2021). "Why and how the brain weights contributions from a mixture of experts". In: *Neuroscience & Biobehavioral Reviews*.

Olsson, Andreas and Elizabeth A Phelps (2007). "Social learning of fear". In: *Nature neuroscience* 10.9, pp. 1095–1102.

Padoa-Schioppa, Camillo and John A Assad (2006). "Neurons in the orbitofrontal cortex encode economic value". In: *Nature* 441.7090, pp. 223–226.

Padoa-Schioppa, Camillo and Xinying Cai (2011). "Orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives". In: *Annals of the New York Academy of Sciences* 1239, p. 130.

Pauli, Wolfgang M, Giovanni Gentile, et al. (2019). "Evidence for model-based encoding of Pavlovian contingencies in the human brain". In: *Nature Communications* 10.1, pp. 1–11.

Pauli, Wolfgang M, Tobias Larsen, et al. (2015). "Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning". In: *Journal of Neuroscience* 35.42, pp. 14220–14233.

Pavlov, P Ivan (1927). "Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex". In: *Annals of Neurosciences* 17.3, p. 136.

Payzan-LeNestour, Élise and Peter Bossaerts (2012). "Do not bet on the unknown versus try to find out more: estimation uncertainty and "unexpected uncertainty" both modulate exploration". In: *Frontiers in Neuroscience* 6, p. 150.

Pickens, Charles L et al. (2003). "Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task". In: *Journal of Neuroscience* 23.35, pp. 11078–11084.

Piray, Payam et al. (2019). "Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies". In: *PLoS Computational Biology* 15.6, e1007043.

Plassmann, Hilke, John P O'Doherty, and Antonio Rangel (2010). "Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making". In: *Journal of neuroscience* 30.32, pp. 10799–10808.

Platt, Michael L and Paul W Glimcher (1999). "Neural correlates of decision variables in parietal cortex". In: *Nature* 400.6741, pp. 233–238.

Pool, Eva R et al. (2019). "Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in humans". In: *Nature human behaviour* 3.3, pp. 284–296.

Poulos, Constantine X, Riley E Hinson, and Shepard Siegel (1981). "The role of Pavlovian processes in drug tolerance and dependence: Implications for treatment". In: *Addictive Behaviors* 6.3, pp. 205–211.

Pouratian, Nader et al. (2012). "Deep brain stimulation for the treatment of Parkinson's disease: efficacy and safety". In: *Degenerative neurological and neuromuscular disease* 2, p. 107.

Pratt, Wayne E and Sheri JY Mizumori (2001). "Neurons in rat medial prefrontal cortex show anticipatory rate changes to predictable differential rewards in a spatial memory task". In: *Behavioural Brain Research* 123.2, pp. 165–183.

Prévost, Charlotte, Mimi Liljeholm, et al. (2012). "Neural correlates of specific and general Pavlovian-to-Instrumental Transfer within human amygdalar subregions: a high-resolution fMRI study". In: *Journal of Neuroscience* 32.24, pp. 8383–8390.

Prévost, Charlotte, Daniel McNamee, et al. (2013). "Evidence for model-based computations in the human amygdala during Pavlovian conditioning". In: *PLoS Computational Biology* 9.2, e1002918.

Price, Joseph L (1999). "Prefrontal cortical networks related to visceral function and mood". In: *Annals of the New York Academy of Sciences* 877.1, pp. 383–396.

Redish, A David and Joshua A Gordon (2016). *Computational psychiatry: New perspectives on mental illness*. Vol. 20. MIT Press.

Redish, A David, Steve Jensen, and Adam Johnson (2008). "Addiction as vulnerabilities in the decision process". In: *Behavioral and Brain Sciences* 31.4, pp. 461–487.

Rescorla, Robert (1988). "Pavlovian conditioning: It's not what you think it is." In: *American psychologist* 43.3, p. 151.

Rescorla, Robert and Allan Wagner (Jan. 1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement". In: vol. Vol. 2. Appleton-Century-Crofts, pp. 64–99.

Rescorla, Robert A (1988). "Behavioral studies of Pavlovian conditioning". In: *Annual review of neuroscience* 11.1, pp. 329–352.

Rich, Erin L and Joni D Wallis (2016). "Decoding subjective decisions from orbitofrontal cortex". In: *Nature Neuroscience* 19.7, pp. 973–980.

Rigoux, Lionel et al. (2014). "Bayesian model selection for group studies—revisited". In: *Neuroimage* 84, pp. 971–985.

Rizzolatti, Giacomo et al. (1996). "Premotor cortex and the recognition of motor actions". In: *Cognitive Brain Research* 3.2, pp. 131–141.

Robinson, Mike JF and Kent C Berridge (2013). "Instant transformation of learned repulsion into motivational "wanting"". In: *Current Biology* 23.4, pp. 282–289.

Rogers, Timothy T. and James L. McClelland (2014). "Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition". In: *Cognitive Science* 38.6, pp. 1024–1077. DOI: 10.1111/cogs.12148. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12148.

Rorie, Alan E et al. (2010). "Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey". In: *PloS One* 5.2, e9308.

Rosene, Douglas L and Gary W Van Hoesen (1977). "Hippocampal efferents reach widespread areas of cerebral cortex and amygdala in the rhesus monkey". In: *Science* 198.4314, pp. 315–317.

Rudebeck, Peter H and Elisabeth A Murray (2011). "Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior". In: *Journal of Neuroscience* 31.29, pp. 10569–10578.

Rudebeck, Peter H, Joshua A Ripple, et al. (2017). "Amygdala contributions to stimulus–reward encoding in the macaque medial and orbital frontal cortex during learning". In: *Journal of Neuroscience* 37.8, pp. 2186–2202.

Rumelhart, David E., James L. McClelland, and CORPORATE PDP Research Group, eds. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press. ISBN: 026268053X.

Rushworth, Matthew FS et al. (2012). "Valuation and decision-making in frontal cortex: one or many serial or parallel systems?" In: *Current Opinion in Neurobiology* 22.6, pp. 946–955.

Rutishauser, Ueli, Adam N Mamelak, and Ralph Adolphs (2015). "The primate amygdala in social perception–insights from electrophysiological recordings and stimulation". In: *Trends in Neurosciences* 38.5, pp. 295–306.

Rutishauser, Ueli, Erin M Schuman, and Adam N Mamelak (2006). "Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo". In: *Journal of Neuroscience Methods* 154.1-2, pp. 204–224.

Rutishauser, Ueli, Shengxuan Ye, et al. (2015). "Representation of retrieval confidence by single neurons in the human medial temporal lobe". In: *Nature Neuroscience* 18.7, p. 1041.

Saez, Ignacio et al. (2018). "Encoding of multiple reward-related computations in transient and sustained high-frequency activity in human OFC". In: *Current Biology* 28.18, pp. 2889–2899.

Sallet, Jérôme et al. (2013). "The organization of dorsal frontal cortex in humans and macaques". In: *Journal of Neuroscience* 33.30, pp. 12255–12274.

Salzman, C Daniel and Stefano Fusi (2010). "Emotion, cognition, and mental state representation in amygdala and prefrontal cortex". In: *Annual Review of Neuroscience* 33, pp. 173–202.

Schmitzer-Torbert, N 1 et al. (2005). "Quantitative measures of cluster quality for use in extracellular recordings". In: *Neuroscience* 131.1, pp. 1–11.

Schoenbaum, Geoffrey, Andrea A Chiba, and Michela Gallagher (1998). "Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning". In: *Nature Neuroscience* 1.2, p. 155.

Schoenbaum, Geoffrey, Barry Setlow, et al. (2003). "Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala". In: *Neuron* 39.5, pp. 855–867.

Schreiber, Thomas and Andreas Schmitz (2000). "Surrogate time series". In: *Physica D: Nonlinear Phenomena* 142.3-4, pp. 346–382.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). "A neural substrate of prediction and reward". In: *Science* 275.5306, pp. 1593–1599.

Seymour, Ben et al. (2004). "Temporal difference models describe higher-order learning in humans". In: *Nature* 429.6992, pp. 664–667.

Sharpe, Melissa J, Chun Yun Chang, et al. (2017). "Dopamine transients are sufficient and necessary for acquisition of model-based associations". In: *Nature Neuroscience* 20.5, pp. 735–742.

Sharpe, Melissa J and Geoffrey Schoenbaum (2016). "Back to basics: Making predictions in the orbitofrontal–amygdala circuit". In: *Neurobiology of learning and memory* 131, pp. 201–206.

Siddle, David AT, Bob Remington, and Muriel Churchill (1985). "Effects of conditioned stimulus preexposure on human electrodermal conditioning". In: *Biological Psychology* 20.2, pp. 113–127.

Siegel, Shepard and Lorraine G Allan (1996). "The widespread influence of the Rescorla-Wagner model". In: *Psychonomic Bulletin & Review* 3.3, pp. 314–321.

Silver, David et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587, pp. 484–489.

Stefanini, Fabio et al. (2019). "A distributed neural code in the dentate gyrus and CA1". In: *bioRxiv*, p. 292953.

Strait, Caleb E, Tommy C Blanchard, and Benjamin Y Hayden (2014). "Reward value comparison via mutual inhibition in ventromedial prefrontal cortex". In: *Neuron* 82.6, pp. 1357–1366.

Sutton, Richard S (1988). "Learning to predict by the methods of temporal differences". In: *Machine Learning* 3.1, pp. 9–44.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, USA.

Suzuki, Shinsuke, Logan Cross, and John P O'Doherty (2017). "Elucidating the underlying components of food valuation in the human orbitofrontal cortex". In: *Nature neuroscience* 20.12, pp. 1780–1786.

Talmi, Deborah et al. (2008). "Human Pavlovian–instrumental transfer". In: *Journal of Neuroscience* 28.2, pp. 360–368.

Taubert, Jessica et al. (2018). "Amygdala lesions eliminate viewing preferences for faces in rhesus monkeys". In: *Proceedings of the National Academy of Sciences* 115.31, pp. 8043–8048.

Thorndike, Edward L (1898). "Animal intelligence: An experimental study of the associative processes in animals." In: *The Psychological Review: Monograph Supplements* 2.4, p. i.

Tobler, Philippe N et al. (2006). "Human neural learning depends on reward prediction errors in the blocking paradigm". In: *Journal of Neurophysiology*.

Tolman, Edward C (1948). "Cognitive maps in rats and men." In: *Psychological Review* 55.4, p. 189.

Trudel, Nadescha et al. (2021). "Polarity of uncertainty representation during exploration and exploitation in ventromedial prefrontal cortex". In: *Nature Human Behaviour* 5.1, pp. 83–98.

Tully, Tim and William G Quinn (1985). "Classical conditioning and retention in normal and mutantDrosophila melanogaster". In: *Journal of Comparative Physiology A* 157.2, pp. 263–277.

Tyszka, J Michael and Wolfgang M Pauli (2016). "In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template". In: *Human Brain Mapping* 37.11, pp. 3979–3998.

Van Den Bos, Ruud, Jolle Jolles, and Judith Homberg (2013). "Social modulation of decision-making: a cross-species review". In: *Frontiers in Human Neuroscience* 7, p. 301.

Vassena, Eliana et al. (2014). "Dissociating contributions of ACC and vmPFC in reward prediction, outcome, and choice". In: *Neuropsychologia* 59, pp. 112–123.

Vogt, Brent A and Deepak N Pandya (1987). "Cingulate cortex of the rhesus monkey: II. Cortical afferents". In: *Journal of Comparative Neurology* 262.2, pp. 271–289.

Voon, Valerie, Kwangyeol Baek, et al. (2015). "Motivation and value influences in the relative balance of goal-directed and habitual behaviours in obsessive-compulsive disorder". In: *Translational psychiatry* 5.11, e670–e670.

Voon, Valerie, Andrea Reiter, et al. (2017). "Model-based control in dimensional psychiatry". In: *Biological psychiatry* 82.6, pp. 391–400.

Wallis, Joni D (2007). "Orbitofrontal cortex and its contribution to decision-making". In: *Annu. Rev. Neurosci.* 30, pp. 31–56.

Wallis, Joni D and Earl K Miller (2003). "Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task". In: *European Journal of Neuroscience* 18.7, pp. 2069–2081.

Walters, Edgar T, Thomas J Carew, and Eric R Kandel (1981). "Associative learning in Aplysia: Evidence for conditioned fear in an invertebrate". In: *Science* 211.4481, pp. 504–506.

Walton, Mark E, Timothy EJ Behrens, et al. (2010). "Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning". In: *Neuron* 65.6, pp. 927–939.

Walton, Mark E, Joseph T Devlin, and Matthew FS Rushworth (2004). "Interactions between decision making and performance monitoring within prefrontal cortex". In: *Nature Neuroscience* 7.11, pp. 1259–1265.

Wang, Shuo, Adam N Mamelak, et al. (2019). "Abstract goal representation in visual search by neurons in the human pre-supplementary motor area". In: *Brain* 142.11, pp. 3530–3549.

Wang, Shuo, Rongjun Yu, et al. (2017). "The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity". In: *Nature Communications* 8, p. 14821.

White, Kate and Graham CL Davey (1989). "Sensory preconditioning and UCS inflation in human 'fear'conditioning". In: *Behaviour research and therapy* 27.2, pp. 161–166.

Wilson, Robert C et al. (2014). "Humans use directed and random exploration to solve the explore–exploit dilemma." In: *Journal of Experimental Psychology: General* 143.6, p. 2074.

Wittmann, Bianca C et al. (2008). "Striatal activity underlies novelty-based choice in humans". In: *Neuron* 58.6, pp. 967–973.

Wunderlich, Klaus, Antonio Rangel, and John P O'Doherty (2009). "Neural computations underlying action-based decision making in the human brain". In: *Proceedings of the National Academy of Sciences* 106.40, pp. 17199–17204.

Yamins, Daniel LK and James J DiCarlo (2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3, pp. 356–365.

Yanike, Marianna and Vincent P Ferrera (2014). "Representation of outcome risk and action in the anterior caudate nucleus". In: *Journal of Neuroscience* 34.9, pp. 3279–3290.

Yartsev, Michael M et al. (2018). "Causal contribution and dynamical encoding in the striatum during evidence accumulation". In: *Elife* 7, e34929.

Yin, Henry H et al. (2005). "The role of the dorsomedial striatum in instrumental conditioning". In: *European Journal of Neuroscience* 22.2, pp. 513–523.