Spring 5-15-2022

Salience Coding in the Basal Forebrain and the Heterogeneous Underpinnings Underlying Novelty Computations

Kaining Zhang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Biomedical Engineering

Dissertation Examination Committee:
Ilya Monosov, Chair
ShiNung Ching
Alexxai Kravitz
Dan Moran
Jeffrey Zacks

Salience Coding in the Basal Forebrain and the Heterogeneous Underpinnings Underlying Novelty Computations

by
Kaining Zhang

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2022
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgements

This dissertation was completed at Washington University in St. Louis between Sept. 2016 and May 2022. It was made possible by the support from the Department of Biomedical Engineering Graduate Program and the Department of Neuroscience.

This dissertation was supported by the National Institute of Mental Health under award number R01MH110594 and R01MH116937, the McKnight Foundation award to Ilya Monosov, and the Defense Advanced Research Projects Agency (DARPA) Biological Technologies Office (BTO) ElectRx program through the CMO grant/contract no. HR0011-16-2-0022.

This dissertation was also edited by the scientific editing network InPrint at Washington University in St. Louis.

During my Ph.D. period, I received tremendous support and help from many people.

First of all, I thank my mentor, Ilya Monosov. He has a lot of passion for studying neuroscience and has an excellent scientific sense. His passion always infects me and motivates me to work hard, and his good scientific sense gives me confidence in my projects. He is a role model for me as a neuroscientist in many aspects. He also cares about his students. He can see the potentials in the students and inspire them to become better. I feel thankful that I have such a great mentor at the beginning of my academic career.

I give special thanks to the senior scientist in our Lab, Ethan Bromberg-Martin. He is one of the smartest and most knowledgeable people I have encountered. He is also kind and generous to help other people. He can always bring up ideas or solutions that refresh my mind as we talk,

which inspires me to catch up to him. I see him as both my mentor and friend. I am fortunate that I have the chance to encounter him and work with him.

I thank Kim Kocher, our lab manager. She provided tremendously good care to our lab animals, trained me to do animal experiments, and even helped us collect experimental data. I would not be able to conduct my projects successfully without her.

I am delighted to work with such a great team in the Monosov lab: Takaya Ogaswara, Ethan Bromberg-Martin, J. Kael White, Julia Pai, Ahmad Jezzini, Yang-Yang Feng, Kim Kocher, Fatih Sogukpinar, Shashank Anand, Jamie Moffa, Charles Chen, Kei Ogasawara, Christopher Teng, Mary Carter and all other members in Monosov Lab past and present. They bring a lot of fun and interesting scientific discussions to the lab and form an intellectual and welcoming space for doing research. They are my collaborators and friends.

I thank my committee members ShiNung Ching, Alexxai (Lex) Kravitz, Dan Moran, Jeffrey Zacks. They guided my dissertation work by giving me valuable comments and feedback. I also thank the professors who have taught me in classes. ShiNung Ching and Barani Raman introduced me to the field of computational neuroscience. Dan Moran taught me the knowledge of electrophysiology, which turned out to be very useful when I did electrophysiological recordings. Rachel Roberts and Aliakbar Daemi gave me detours from neuroscience and showed me the beauty of topology and geometry. Heinz Schaettler taught me probability and optimal control theory. These classes in mathematics provided me with good mathematical background and trained me to think more systematically and rigorously. Many more people have taught me knowledge in different fields and gave me inspirations, including Renato Feres, Sanjoy Baruah, Jonathan Silva, Yakov Berchenko-Kogan, Cathy Raymond, Jin-Yu Shao, Tim Holy, Krikor

Dikranian, Camillo Padoa-Schioppa, Larry Snyder, Randall Hoven, Wilson Ray, Matthew MacEwan, Ying Yan, and Youchun Zeng.

Another important gain in my Ph.D. life is that I developed many valuable friendships with people inside and outside the school community. We hiked, traveled, climbed, skied, played board games, and chatted about experiences and sentiments in our lives. I thank these friends, especially Hao Chen, Weikang Shi, Binxu Wang, Zhikai Liu, and Ethan Bromberg-Martin, for their tremendous support and encouragement; also including Xiaodan Wang, Lifei Zhu, Yang-yang Feng, Fatih Sogukpinar, Pingchuan Ma, Yuechen Shen, Zihao Chen, Yijian Xiang, Chenxi Yin, Joshua Falconer, Brewer Kight, Ilia Katritch, Chantal Chen, Pratyush Ramakrishna, Cindy Tu, Ying Xiang, Bryant Moy, Hark Lee, Wentao Han, Geng Wang, Xue Yang, Keran Yang, Yifan Xu, Aelita Zhu, and Manning Zhang. I value these friendships and thank you all for providing such a friendly and warm space where I can thrive academically and personally.

Lastly, I thank my mom and my grandparents. You have been giving me unconditional love and robust support. The more I grew up, the further I realized how difficult and challenging it is to raise a child up, and I cannot thank you enough. My mom's encouragement to my curiosity in childhood made me deeply attracted to natural sciences, which eventually led me to the path of biomedical engineering and neuroscience today.

Kaining Zhang

Washington University in St. Louis

May 2022

ABSTRACT OF THE DISSERTATION

Salience Coding in the Basal Forebrain and

the Heterogeneous Underpinnings Underlying Novelty Computations

by

Kaining Zhang

Doctor of Philosophy in Biomedical Engineering

Washington University in St. Louis, 2022

Professor Ilya Monosov, Chair

Humans and animals are consistently learning from the environment by interacting with it and getting feedback from their actions. In the environment, some objects are more important than others, because they are associated with reward, uncertainty, surprise, or novelty etc. These objects are salient to the animal. Salient objects attract attention and orientation, increase arousal, facilitate learning and memory, and affect reinforcement learning and credit assignment. However, the neural basis to support these effects is still not fully understood.

We first studied how the basal forebrain, one of the principal sources of modulation of the neocortex, encodes salience events. We found two types of neurons that process salient events in distinct manners: one with phasic burst activity to cues predicting salient events and one with ramping activity anticipating such events. Bursting neurons respond to reward itself and cues that predict the magnitude, probability, and timing of reward. However, they do not have a selective response to reward omission. Thus, bursting neurons signal surprise associated with external events, which is different from the reward prediction error signaled by the midbrain dopamine neurons. Furthermore, they discriminate fully expected novel visual objects from familiar objects and respond to object-sequence violations. In contrast, ramping neurons predict the timing of

many salient, novel, and surprising events. Their ramping activity is highly sensitive to the subjects' confidence in event timing and on average encodes the subjects' surprise after unexpected events occur. These data suggest that the primate BF contains mechanisms to anticipate the timing of a diverse set of salient external events (via tonic ramping activity) and to rapidly deploy cognitive resources when these events occur (via phasic bursting activity).

Then we sailed out to study one special salience signal – Novelty. The basal forebrain responds to novelty, but the neuronal mechanisms of novelty detection remain unclear. Prominent theories propose that novelty is either derived from the computation of recency or is a form of sensory surprise. Here, we used high-channel electrophysiology in primates to show that, in many prefrontal, temporal, and subcortical brain areas, object novelty sensitivity is related to both computations of recency (the sensitivity to how long ago a stimulus was experienced) and sensory surprise (violation of predictions about incoming sensory information). Also, we studied neuronal novelty-to-familiarity transformations during learning across many days and found a diversity of timescales in neurons' learning rates and between-session forgetting rates within and across brain regions that is well suited to support flexible behavior and learning in response to novelty. These findings show that novelty sensitivity arises on multiple timescales across single neurons due to diverse related computations of sensory surprise and recency, and shed light on the logic and computational underpinnings of novelty detection in the primate brain.

Chapter 1: Introduction

1.1 Important objects

Humans and animals are consistently learning from the environment by interacting with it and getting feedback from their actions. In the environment, some objects are more important than others because they are directly associated with rewards or punishment, or they have information, indicate possible changes, or are novel and therefore are important for adaptive learning and memory (Fecteau and Munoz, 2006; Ponzi, 2008; Bromberg-Martin et al., 2010a; Barto et al., 2013; Ghazizadeh et al., 2016a; Zhu et al., 2018; Parr and Friston, 2019).

The first step of processing important objects or events is to identify them as important. It may sound trivial, but a few things happen when someone notices an important object or event. For example, when the fire alarm rings when someone is sleepy, they will immediately wake up (arousal increasing), look at where the alarm is coming from (head and eyes orientating), and start to figure out what is going on with the alarm (attention drawing). Increasing arousal, drawing attention, and eliciting orientation are the typical effects when noticing and responding to important events or objects (Ohman et al., 2001; Fecteau and Munoz, 2006; Ghazizadeh et al., 2016a; Zhu et al., 2018; Parr and Friston, 2019).

Objects that evoke these psychological and behavioral effects are often called "salient". There are many features that can contribute to object salience; for example, how bright or colorful the object is, whether it is moving, and other physical features of the object (physical salience) (Bachman et al., 2020). However, other features also contribute to object salience. For example,

when someone is waiting for an important notice on their phone, such as a text, and their phone vibrates lightly for a second, they may notice immediately and overreact, experiencing a strong physiological response. Objects with the same physical features can have different object salience, depending on the context, prior history, and current internal state (Ghazizadeh et al., 2016a; Bachman et al., 2020).

Ghazizadeh et al. (2016a) systematically studied some ecological features that contribute to object salience, which is related to the animal's past experience with the object rather than the physical features, by analyzing gaze behavior in primates. They found that object salience can be increased by 1) how much reward the object is associated with, 2) the uncertainty of the outcome that the object is associated with, 3) the novelty of the object, and 4) some aversive events, like threats, that the object is associated with. They term the salience that arises from them “ecological salience”.

Furthermore, some objects motivate animals to change their behavior, causing them to approach or avoid the object. For example, delicious food might elicit approach behavior, while the image of a predator might motivate avoidance. These objects elicit motivational salience and are usually associated with values (Bromberg-Martin et al., 2010a; Puglisi-Allegra and Ventura, 2012). In primates, this is particularly clear when monkeys rapidly orient to salient objects but then ultimately avoid choosing them (Jezzini et al., 2021).

How does the brain process object salience, ecological salience, and motivational salience?

In this dissertation, I will first study how the basal forebrain, one of the neuromodulatory hubs that mediates neocortical computations, responds to objects with motivational salience or

ecologically salience. Then I will switch my focus to one attribute of salience – object novelty, and study how the brain computes and processes the novelty signals.

1.2 Basal forebrain (BF) neuromodulatory system

The basal forebrain (BF) contains several nuclei, including nucleus basalis, diagonal band, medial septal, and substantia innominate, and has wide projections to and influences on the neocortex, hippocampus, and amygdala (Mesulam et al., 1983; Baxter and Chiba, 1999; Turchi et al., 2018). The BF is the major hub of the cholinergic neuromodulatory system but also contains GABAergic and glutaminergic neurons. Moreover, the loss of neurons in the BF can predict the development of Alzheimer's disease and Parkinson's disease (Whitehouse et al., 1982; Arendt et al., 1995; Pereira et al., 2020).

One hypothesis of the function of the BF cholinergic system is that it regulates attention. Some animal experiments have shown that lesion or inhibition of cholinergic neurons in the BF increases the animal's reaction time and decrease the accuracy of detecting and discriminating sensory stimuli (Bucci et al., 1998; Chiba et al., 1999; Waite et al., 1999; Pinto et al., 2013). On the other hand, stimulation of the same area can improve task performance (Pinto et al., 2013).

Other studies focus on the BF's role in cortical plasticity and memory formation. Inhibition of the BF could impair animal memory formation (Hasselmo and Schnell, 1994; Gu and Yakel, 2011). However, the findings of the BF influencing memory formation are inconsistent between species. Compared with rodents, memory impairment was more subtle when lesioning primates' BF (Voytko et al., 1994).

In addition, the BF is also known for participating in the control of the sleep-wake cycle. Lesion or inhibition of the BF led to sleep loss in some studies (McGinty and Sterman, 1968; Szymusiak and McGinty, 1986) and had complex influences on the delta or theta electroencephalogram (EEG) bands (Brown et al., 2012). Furthermore, the cholinergic neurons, some GABAergic and glutaminergic neurons in the BF burst during waking and rapid eye movement (REM) sleep (Lee et al., 2005; Xu et al., 2015).

the majority of the neurons in the BF are cholinergic, but it also contains some GABAergic and glutaminergic projection neurons (Mesulam et al., 1983), which have similar projections as the cholinergic neurons (Gritti et al., 1998). The functions of GABAergic and glutaminergic neurons are less studied. In the studies that include those neurons, the results vary from each other. In general, studies from different groups have demonstrated that the noncholinergic neurons also regulate the sleep-wake cycle, attention, and cortical plasticity (Lin and Nicolelis, 2008; Avila and Lin, 2014; Hangya et al., 2015; Xu et al., 2015).

Electrophysiology and optogenetic methods have been used to record single neuron activity in the BF. A series of rodent studies found that a subset of BF neurons has a phasic bursting activity pattern, and they encode both reward and aversive events in the same direction by enhancing their activity (Lin and Nicolelis, 2008; Avila and Lin, 2014; Hangya et al., 2015). In other words, they encode motivational salience. They also discovered another subset of neurons with a tonic activity pattern, whose activities are correlated with behavioral reaction time. In addition, studies in primates demonstrate that the primate tonic BF neurons respond to reward and uncertain events by ramping up their activities to the possible reward delivery time (Monosov et al., 2015; Ledbetter et al., 2016b). However, there was no report of the phasic bursting neurons in the

primate BF before our experiment. Furthermore, there is a lack of study in the BF's coding of the other ecological salience signals like novelty and sensory surprise, besides the motivational salience signals, like reward and punishment.

In addition, the BF system sometimes is compared with the midbrain dopaminergic neuromodulatory system. Midbrain dopamine (DA) neurons are mostly known for their encoding of reward prediction error (RPE), which is defined as the differences between received reward and the prediction, and unsigned RPE, which is the absolute value of the RPE (Schultz et al., 1997; Matsumoto and Hikosaka, 2007; Bromberg-Martin and Hikosaka, 2011; Schultz, 2016). For DA neurons that encode unsigned RPE, they respond to both appetitive and aversive events by enhancing their firing rates, which is similar to the activity of BF phasic neurons. Studying the similarity and difference of the coding and function between different neuromodulatory systems is still a major topic in neuroscience (Avery and Krichmar, 2017).

In Chapter 2, we report that the primate BF contains at least two functional subtypes of neurons that often process salient events in distinct manners: one with phasic burst activity to cues predicting salient events and one with ramping activity anticipating such events. Bursting neurons respond to cues that convey predictions about the magnitude, probability, and timing of primary reinforcements. However, they do not have a selective response to reinforcement omission (the unexpected absence of an event). Thus, bursting neurons do not convey reward prediction errors but instead signal surprise associated with external events. Indeed, they are not limited to processing primary reinforcement: they discriminate fully expected novel visual objects from familiar objects and respond to object-sequence violations. In contrast, ramping neurons predict the timing of many salient, novel, and surprising events. Their ramping activity

is highly sensitive to the subjects' confidence in event timing and encodes the subjects' surprise after unexpected events occur. These data suggest that the primate BF contains mechanisms to anticipate the timing of a diverse set of important external events (via ramping activity) and to rapidly deploy cognitive resources when these events occur (via short-latency bursting).

1.3 Novelty signals and computations in the brain

After studying how the basal forebrain encodes salient events, we shift our focus to one particular salience signal – novelty.

A novel object is an object presented for the first time to the animal, and some features of the object deviate from the animal's previous experience (Markou and Singh, 2003; Barto et al., 2013). Detecting, pursuing, exploring, and memorizing novel objects are the crucial steps for humans and animals to learn from the environment (Barto et al., 2013; Jaegle et al., 2019; Ogasawara et al., 2022). Studies have found novelty signals in multiple brain areas (Berns et al., 1997; Ranganath and Rainer, 2003; Yamaguchi et al., 2004), but how novelty signals are computed in the brain is still not fully understood.

A study in 1965 firstly reported that the brain signals novelty in the EEG (Sutton et al., 1965; Ranganath and Rainer, 2003). There is a specific signal related to novel events in EEG – The Novelty P3. It is an event-related potential (ERP) that is elicited by novel or unexpected events peaking around 300ms after the event and is presumably related to attention. (Sutton et al., 1965; Cycowicz et al., 2001; Friedman et al., 2001; Polich, 2007)

Novelty P3 has different types and distributions across the scalp. They can be categorized into two main types: P3a and P3b. P3a is more frontal and is elicited by novel or surprising events;

P3b is more parietal and is elicited especially by novel or surprising task-relevant events. Novelty P3 signals habituate to repeating novel events quickly, within the first few repetitions in a session. However, they also show heterogeneity. The signal in the frontal scalp decays faster than that in the parietal scalp (Friedman et al., 2001). In Chapter 3, we show recordings from single neurons in different brain regions, and the results also support the heterogeneity in novelty habituation.

Then, many MRI and PET studies of novelty found multiple brain areas responding to novel events. (Tulving et al., 1996; Berns et al., 1997; Kiehl et al., 2001; Ranganath and Rainer, 2003; Yamaguchi et al., 2004; Hawco and Lepage, 2014). Most studies found that the hippocampus, some areas of the temporal cortex, and the frontal cortex have novelty responses, though results varied slightly. Some studies, in addition, reported novelty responses in striatum and cingulate cortex (Berns et al., 1997), occipital cortex and parietal cortex (Tulving et al., 1996; Hawco and Lepage, 2014), and insula (Kiehl et al., 2001).

Novel events have multiple effects on the animal. They can provoke orientation behavior, increase arousal and attention, which are shared with other salient events (Bradley, 2009; Schomaker and Meeter, 2015). In addition, novelty events also increase learning (Tulving and Kroll, 1995; Hasselmo et al., 1996; Meeter et al., 2004). One hypothetical neural mechanism is through controlling the level of acetylcholine in the hippocampus, which is further controlled by the BF neuromodulatory system (Hasselmo et al., 1996; Meeter et al., 2004; Hasselmo and Sarter, 2011; Zaborszky et al., 2018).

Furthermore, most animals seek novel events. One theory of how the brain produces novelty-seeking is that novel objects provoke intrinsic reward in the brain, and thus the reward circuit

guides the behavior of exploring novel objects (Kakade and Dayan, 2002; Jaegle et al., 2019). However, more recent experiments show that reward-seeking and novelty-seeking circuits can differ (Foley et al., 2014; Ogasawara et al., 2022). Ogasawara et al. (2022) demonstrated that the circuit regulating novelty-seeking includes zona incerta (ZI) and anterior ventral medial temporal cortex (AVMTC), but not lateral habenula (LHb) and substantia nigra (SN) which are the areas traditionally associated with reward-seeking.

Novel objects have multiple psychological and behavioral effects and are presumably regulated by different circuits (Schomaker and Meeter, 2015). The next question is, how does the brain detect novel objects from familiar ones? Furthermore, for different circuits that regulate different effects of novelty, do they share the same computation of novelty signal?

Many possible mechanisms of novelty detection at the circuit level have been proposed. Bogacz et al. (2001b) proposed that novelty detection in the primate entorhinal cortex is presumably implemented by a feedforward neural network deriving from the Hopfield network (Hopfield, 1982). In fruit flies, Sanjoy Dasgupta et al. (2018) proposed that in the mushroom body, novelty detection is presumably implemented by a method deriving from the Bloom filter (Bloom, 1970). Furthermore, Danil Tyulmankov et al. (2022) used the meta-learning method to train feedforward networks with the Hebbian or anti-Hebbian rules to generate some biologically possible novelty detection models. These circuit-level mechanisms store the information of familiar objects in the synaptic connections, and the connections keep changing to accommodate new objects.

In all these circuit-level novelty detection models, the objective goal is purely to differentiate objects seen for the first time vs. those seen many times. However, novel objects also have other

properties: They are surprising, and no similar objects have been seen recently (Markou and Singh, 2003; Barto et al., 2013; Pimentel et al., 2014). Accordingly, other mechanisms have been proposed at a higher cognitive level.

Novelty is closely related to surprise, which can be defined as the mismatch of the incoming sensory stimuli and the prediction. Some papers also refer to the surprising objects in their behavior procedures as the contextual novelty (Ranganath and Rainer, 2003; Nyberg, 2005). Novel objects are usually surprising, but surprising objects are not necessarily novel (Strange and Dolan, 2001; Barto et al., 2013). An MRI study has demonstrated that the hippocampus and some areas in the frontal cortex respond to both novel and surprising events (Strange and Dolan, 2001). One study proposed a theory that the novelty response in the hippocampus is actually generated by a mechanism of mismatch, aka, surprise (Kumaran and Maguire, 2007b).

Novelty is also closely related to recency. The degree of novelty of an object is related to how recently the animal has seen the same or similar object. In the medial/inferior temporal cortex, some neurons respond differently to familiar objects presented recently vs. not recently (recency response)(Xiang and Brown, 1998). The neural recency response is also observed in the primary visual cortex, which is an initial area to process visual information, and the finding can date back to a study in cat in 1960s (Hubel and Wiesel, 1962; Dragoi et al., 2000). A theory has been proposed that the recency responses can be generated through local synaptic change (Vogels, 2016) and can support the function of novelty detection (Bogacz et al., 2001a).

To summarize, according to the hypothetical mechanisms of the novelty detection that have been proposed, there are four testable hypotheses: 1) Novelty computation could arise with surprise computation which computes novelty as a form of sensory surprise; 2) Novelty computation

could arise with recency computation which computes novelty as recency and/or repetition effect; 3) Novelty computation could arise with both surprise and recency computations; 4) Novelty computation could arise independently of surprise or recency, which purely serves the goal of differentiating objects that have been seen for the first time vs. many times.

There is no study that compares all these hypotheses simultaneously; thus, in Chapter 3, we tested all these hypotheses for the first time. In addition, we used high-channel-count electrode arrays to record novelty responses in multiple brain areas, which have more accuracy both spatially (up to single neuron) and temporally (up to single spike) compared with previous fMRI, PET, and EEG studies.

We found that at the single neuron level, the computation of novelty depends on both sensory surprise and recency. This dependency is observed both within brain areas and across brain areas. However, different brain areas do not share precisely the same computation.

We also investigated how single neurons adapt as novel objects gradually become familiar. We presented the same novel objects repeatedly to the animals for multiple days. Neurons that are excited by novel objects gradually decrease their firing rate as the repeating novel objects' presentation number increases. We measured the learning rate of the neuron population. The learning rate drops as the repeating novel objects' presentation number increases. In addition, different neurons also show heterogeneous timescales in their adaptation to repeating novel objects, and this heterogeneity also exists among different brain areas.

1.4 Summary

Salience signals play an important role in guiding our behavior. Salient objects include rewards, punishments, cues indicating them, and surprising, uncertain, and novel objects. They make the animal orient towards them, attract attention, and improve arousal. Cumulative evidence has demonstrated that the neuromodulatory systems, especially the basal forebrain neuromodulatory system, regulate many of the behavioral and psychological effects of salience.

In Chapter 2, we studied how the basal forebrain encodes different kinds of salient events. We recorded two subsets of the basal forebrain neurons in the primate brain and studied their activities and tunings to different events with motivational and ecological salience.

Novelty is a special type of salience signal. In Chapter 3, we studied the underpinnings that influence the novelty computations in the brain. We recorded neurons from multiple brain areas and discovered that the computation of novelty is supported by the computations of surprise and recency. We further investigated how the brain goes through the novelty-familiarity transformation when presenting the repeating novel objects.

Chapter 2: Novelty, salience, and surprise timing are signaled by neurons in the basal forebrain [1]

The basal forebrain (BF) is a principal source of modulation of the neocortex and is thought to regulate cognitive functions such as attention, motivation, and learning by broadcasting information about salience. However, events can be salient for multiple reasons - including novelty, surprise, or reward prediction errors - and to date, precisely which salience-related information the BF broadcasts is unclear. Here, we report that the primate BF contains at least two types of neurons that often process salient events in distinct manners: one with phasic burst activity to cues predicting salient events and one with ramping activity anticipating such events. Bursting neurons respond to cues that convey predictions about the magnitude, probability, and timing of primary reinforcements. They also burst to the primary reinforcement itself, particularly when it is unexpected. However, they do not have a selective response to reinforcement omission (the unexpected absence of an event). Thus, bursting neurons do not convey value-prediction errors but do signal surprise associated with external events. Indeed, they are not limited to processing primary reinforcement: they discriminate fully expected novel visual objects from familiar objects and respond to object-sequence violations. In contrast, ramping neurons predict the timing of many salient, novel, and surprising events. Their ramping

[1] This chapter is adapted from a published paper by Kaining Zhang, Charles D. Chen and Ilya E. Monosov: "Novelty, Salience, and Surprise Timing Are Signaled by Neurons in the Basal Forebrain.", Current Biology 29.1 (2019): 134-142.

activity is highly sensitive to the subjects' confidence in event timing and on average encodes the subjects' surprise after unexpected events occur. These data suggest that the primate BF contains mechanisms to anticipate the timing of a diverse set of important external events (via ramping activity) and to rapidly deploy cognitive resources when these events occur (via short latency bursting).

2.1 Introduction

The basal forebrain (BF) is a principal source of modulation of the neocortex (Mesulam et al., 1983; Everitt and Robbins, 1997; Baxter and Chiba, 1999; Monosov et al., 2015; Zaborszky et al., 2015; Turchi et al., 2018) and is thought to regulate cognitive functions such as attention, motivation, and learning by broadcasting information about salience (Richardson and DeLong, 1990; Wilson and Rolls, 1990; Fukuda et al., 1993; Voytko, 1996; Masuda et al., 1997; Chudasama et al., 2004; Wilson and Ma, 2004; Pinto et al., 2013; Avila and Lin, 2014; Peck and Salzman, 2014; Hangya et al., 2015; Lin et al., 2015; Raver and Lin, 2015). However, events can be salient for multiple reasons - such as novelty, surprise, or reward prediction errors (Hayden et al., 2011; Preuschoff et al., 2011; Wallis and Rich, 2011; Wang and Mitchell, 2011; Barto et al., 2013) - and to date, precisely which salience-related information the BF broadcasts is unclear.

Previous work suggests that two prominent neuronal activation patterns in the BF support its mediation of cognitive functions in response to salient events: phasic bursting (Lin and Nicolelis, 2008; Hangya et al., 2015), which has been identified in the brains of rodents, and tonic activations (Hangya et al., 2015; Monosov et al., 2015), which in monkeys are often seen in neurons that also ramp to the time of delivery of uncertain or noxious outcomes (Monosov et al., 2015). To date, it remains unclear how these neuronal activations signal surprise and/or novelty

and how their surprise-related responses relate to errors in estimates of state values, referred to as reward prediction errors (RPEs). Therefore, how bursting and ramping BF activations contribute to cognitive functions remains poorly understood. Here, we assessed whether prediction-related phasic bursting and ramping activity occur in distinct groups of neurons and tested whether and how the BF represents prediction errors, surprise, value, novelty, and timing.

2.2 Materials and Methods

2.2.1 Experimental Model

Six adult sexually mature male rhesus monkeys (monkeys B, R, Z, W, H, and P; ages: 7-10 years old) were used for recording experiments. All procedures conform to the Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee at Washington University (monkeys B, R, W, and Z) and the National Eye Institute (monkeys P and H).

2.2.2 Data acquisition

All monkeys underwent surgery under general anesthesia. For each monkey, a plastic head holder and recording chamber were fixed to the skull under general anesthesia and sterile conditions. Chambers were tilted laterally from midline by 35 degrees and aimed at the basal forebrain and anterior portion of striatum. After the monkeys recovered from surgery, they participated in behavioral and neurophysiological experiments.

While the monkeys participated in behavioral procedures we recorded single neurons in the basal forebrain. The recording sites were determined with 1 mm-spacing grid system and with the aid of MR images (3T) obtained along the direction of the recording chamber. This MRI-based

estimation of neuron recording locations was aided by custom-built software. Single-unit recording was performed using glass-coated electrodes (Alpha Omega). During each recording session, an electrode was inserted into the brain through a stainless-steel guide tube and advanced by an oil-driven micromanipulator (MO-97A, Narishige). Signal acquisition (including amplification and filtering) was performed using Alpha Omega 44 kHz SNR system. Action potential waveforms were identified online by multiple time-amplitude windows with an additional template matching algorithm (Alpha-Omega).

Neuronal recordings were restricted to single well-isolated neurons in the basal forebrain that displayed task related ramping or phasic-bursting activity following the presentation of the task conditioned stimuli in the Probability Amount procedure. The ventral pallidum (defined using anatomical criteria and previous electrophysiological criteria, such as high and irregular firing rate) was not part of this study. The locations of the BF recordings are detailed in Supplemental Figure 2.2. Reconstruction procedures were detailed previously (Daye et al., 2013).

Eye position was obtained with an infrared video camera (Eyelink, SR Research). Behavioral events and visual stimuli were controlled by MATLAB (Mathworks, Natick, MA) with Psychophysics Toolbox extensions. Juice, used as reward, was delivered with a solenoid delivery reward system (CRIST Instruments). Juice-related licking was measured and quantified using previously described methods. Airpuffs were delivered through a narrow tube placed ~6-8cm from the monkey's face.

2.2.3 Behavioral tasks

Probability Amount procedure

To study (1) neuronal representations of reward probability and amount, and (2) delivery-related responses following uncertain predictions, we trained monkeys on a Pavlovian conditioning procedure. Pavlovian conditioning was used to avoid fluctuations in reward rate across trials or fluctuations in outcome timing within single trials (related to action performance) which theoretically may affect outcome prediction error signals (Apicella et al., 2011).

The Pavlovian conditioning procedure contained two blocks of trials: a reward-probability block and a reward-amount block. Each trial started with the presentation of a green trial-start cue at the center. The monkeys had to maintain fixation on this trial-start cue for 1 s; then the trial start cue disappeared and one of the CSs was presented pseudo randomly. After 2.5 s (for monkeys B, Z, and R) or 1.5 s (monkeys H and P), the CS disappeared, and juice (if scheduled for that trial) was delivered. The longer duration was introduced for monkey B, Z, and R to verify that the ramping activity in the BF reaches maximum at the time of the outcome across different CS durations. The reward-probability block contained five visual fractal object CSs associated with five probabilistic reward predictions (0, 25, 50, 75 and 100% of 0.25 mL of juice). The reward-amount block contained five objects associated with certain reward predictions of varying reward amounts (0.25, 0.1875, 0.125, 0.065 and 0ml). Each block consisted of 20 trials (monkeys B, Z, and R) and 40 trials (monkeys P and H) with fixed proportions of trial types (each of the five CSs appears four times in each block or 8 times in each block, depending on block length). The expected values of the five CSs in the probability block matched the expected values of the five CSs in the amount block. This two-block design removed confounds introduced by risk seeking-related changes in subjective values of the CSs (Monosov and Hikosaka, 2013; White and Monosov, 2016).

Before neuronal recordings began, the monkeys' knowledge of the CSs was confirmed by a choice procedure that was detailed previously (Monosov and Hikosaka, 2013; Monosov et al., 2015). Briefly, in separate experimental sessions, the monkeys' choice preference was tested for the CSs. Each trial started with the presentation of the trial-start cue at the center, and the monkeys had to fixate it. Then two CSs appeared 10 degrees to the left and right. The monkeys had to make a saccade to one of the two CSs within 5 s and fixate it for at least 750 ms. Then, the unchosen CS disappeared, and after a brief delay the outcome (associated with the chosen CS) was delivered, and the chosen CS disappeared. If the monkey failed to fixate one of the CSs, the trial was aborted and all stimuli disappeared. The trials were presented pseudo randomly, so that a block of 180 trials contained all possible combinations of the 10 CSs four times. To verify that the monkeys' knowledge is stable during recording, we also monitored licking behavior and confirmed that it, like the choices, scaled with the expected values of the probability CSs and amount CSs (two separate Spearman's correlations, threshold: $p < 0.05$). The CS epoch responses of the 31 neurons recorded in monkeys H and P were previously analyzed in (Monosov et al., 2015).

Temporal Uncertainty Procedure

To assess how monkeys' BF neurons encoded uncertain predictions about reward timing, monkeys B, R, Z were trained on an additional Pavlovian procedure (Supplemental Figure 2.3). Following a trial start cue fixation period (same as above), one of five CSs were presented. These CSs predicted either (1) a probabilistic delay before a reward with deterministic delivery (reward-timing-uncertain CSs); or (2) a deterministic delay before a reward with 0.5 probability of delivery (reward-probability CS). In trials with one of the four reward-timing-uncertain CSs,

reward was always delivered either 1.5 s after CS onset or 4.5 s after CS onset. Depending on the reward-timing uncertain CS, the reward was delivered at 1.5 s with 0.25, 0.50, 0.75, or 1 probability. In trials with the reward-probability CS, reward was delivered with a delay of 1.5 s after CS onset with 0.50 probability. During, the 0.25, 0.50, and 0.75 CS trials, when reward was not delivered at 1.5 s, the CS remained on the screen until reward was delivered at 4.5 s. During the 0.50 reward probability CS, the CS turned off at the time of the outcome (when reward was either delivered or omitted). The inter-trial-interval ranged from 2 to 6.5 seconds.

The training was verified by monkeys' reward anticipatory licking behavior. The data suggested that they understood the meanings of the CSs and were highly sensitive to the timing and probability of reward (Supplemental Figure 2.3B). First, during the four reward-timing-uncertain CSs, monkeys displayed increased licking behavior before 1.5 s, then a decrease in licking behavior after 1.5 s if the reward was not delivered, then finally an increase in licking behavior to the time of reward at 4.5 s. During reward omissions, in 75% reward trials licking behavior remained higher than 25% and 50% trials, even 0.5 s after the reward was omitted at 1.5 s ($p < 0.01$, rank-sum test, time window 2 s to 2.5 s after the onset of fractal). Also, the mean magnitude of anticipatory licking behavior before possible reward delivery at 1.5 s across all trials increased with the probability of reward delivery at 1.5 s (Spearman's rank correlation, $\rho = 0.38$, $p = < 0.0001$; Supplemental Figure 2.3). These behavioral results indicate that the magnitude and persistence of the monkeys' anticipatory behavior were strongly influenced by reward timing conveyed by the CSs.

Object Sequence Procedure

An object sequence task was used to study how BF neurons encode sensory predictions and object novelty. Monkeys B, R, and Z experienced four distinct sequences of object presentations (S1, S2, S3, S4). The object sequences began following a 0.5 s period of fixation on the trial start cue that appeared in the center of the screen. Each sequence contained 3 familiar objects and 1 novel object. These objects were presented in the center of the screen and occupied ~3 degrees visual angle. The novel object was always presented in second position in the sequence. Therefore, the novel object was surprising because it was never experienced by the monkeys, but its presentation did not deviate from the animals' expectations. Monkeys performed more than 10,000 trials before recordings began. Following sequences S2 and S4, the monkeys performed a reaction-time Delayed Non-matching-to-Sample task (DNMS). During DNMS, an object that was novel during the presentation of S2 (or S4 if the DNMS trial followed S4) was presented with a novel object that has never been experienced. The objects were presented 10 degrees from the center, to the left and the right of the fixation point. The trial continued until the monkeys fixated the novel object for 0.5 ms to get a reward. The monkeys were never penalized for looking at the previously experienced object. Therefore, the novel objects in S2 or S4 did not have an explicit reward association, but aided the monkey in subsequent DNMS trials. On ~11% of S2 or S4 presentations, the first or the third fractal was replaced by a corresponding fractal from sequences S1 and S3 (in S2 from S1; and in S4 from S3). For example, if the first fractal in S2 was replaced, the first fractal from S1 was always displayed instead. In this way, sequence violations did not alter the relationship of the individual fractals to the timing of reward delivery. We used the probability-amount procedure to identify phasically bursting BF neurons and uncertainty ramping neurons and studied them in the object sequence procedure. All phasic bursting neurons included in Figure 2.5 had greatest responses for 100% reward CSs.

Reward and Novelty Motivated Gaze Task

To test if monkeys are motivated by novelty we trained Monkeys R and Z on a novel saccadic task (Supplemental Figure 2.5) that measured their eagerness to observe a novel visual object. First, a fixation dot appeared in the center of the screen. 0.5 s after the onset of the fixation dot, a visual object fractal appeared 10 degrees to the right or the left of the fixation dot. The monkey was required to continue fixating the dot in the center. After 0.35 s the fixation spot disappeared and the monkey was free to make saccades. Reward was always delivered 3 s after the fractal onset. Therefore, the monkeys' saccadic behavior after the fixation spot disappeared did not affect reward delivery. In this task, the monkeys experienced four different trial types. The first two types of trials contained a novel (type 1) or 1 of 2 familiar (type 2) visual fractal objects. Two additional trial types (3-4) tested whether the monkeys were motivated by the possibility of viewing a novel fractal. In trial type 3, 1 of 2 distinct familiar fractal objects appeared. After the fixation spot disappeared, if the monkey fixated the familiar object, it was immediately replaced by a novel object. In trial type 4, 1 of 2 other distinct familiar objects appeared. If the monkey fixated this object, it was replaced by 1 of 2 other familiar objects. If novelty is salient, we ought to observe faster target acquisition times (duration between the time when the stimulus was presented and when the monkey saccades to its location) in trial type 1 than 2. Also, if novelty exerts motivational effects on saccadic behavior, then we ought to see faster target acquisition times in trial type 3 than 4.

2.2.4 Data processing and statistics

In order to generate spike density functions, spike times were convolved with a Gaussian kernel (σ = 100 ms). Statistical tests were two-tailed. All permutation tests used 10000 shuffles. For all

analyses and figures that included deliveries and omissions of rewards, unless explicitly stated in the text, a neuron was included if it had at least 2 trials for reward delivery and omission.

To cluster the single neurons' average responses in the probability block (Figure 2.1), first we performed principal component analysis (PCA). We then applied Silhouette and Calinski-Harabasz tests to confirm the optimal number of clusters (n = 2). K-means clustering was used to cluster the data based on PCs into 2 clusters (for this, using the first 3 PCs and up to 10 PCs resulted in very similar group membership).

To calculate the latency of reward size coding information (Figure 2.1C) we performed a correlation of firing rate and value in time (in 100 ms bins moving 1 ms steps) for each neuron. For each time bin we calculated the p value of the Spearman's rank correlation of neuron's activity with reward amount in the reward amount block. Reward size coding latency was defined as the first time p was lower than 0.01 (but similar results were obtained at $p < 0.05$). These statistical-latency analyses do not determine the actual latency of information coding per se because they utilize an arbitrary threshold. Instead, they are useful for demonstrating relative latencies across two groups of neurons.

To calculate the baseline rate that was used to derive the latency with which ramping neurons returned to baseline (Figure 2.3A), we picked the time window from 1000 ms to 500 ms before trial start cue appeared and used the average firing rate in this time window as the baseline.

To fit the outcome related activity with exponential functions (Figure 2.3), we first derived spike density functions using overlapping bins of 50 ms (in 20 ms steps). Then we used a least-squares method to fit the data by the function: $A * e^{-\lambda t} + C$, in which λ is the decay rate, representing how fast the firing rate decreases. λ is restrained by the interval (0,0.06). To determine if the

decay rates were significantly different across the different reward-omission conditions we used bootstrapping to calculate the confidence interval of the difference between two decay rates and tested if the 95% confidence interval excluded a difference of zero. Bootstrapping was done by randomly resampling the neurons with replacement (500 times). Each time resampling was done, we obtained a set of decay rates by fitting the neurons' average activity to the function shown above. For Figures 2.2A–2.2D, data from probability-amount and reward timing procedures were pooled (see outcome responses separately in Supplemental Figures 2.1 and 2.3).

In the DNMS object sequence task, reward was delivered as long as the monkey fixated on the novel object for 0.5 s, regardless if he had looked at the other object. To evaluate the monkey's performance, we focused on the primary choice the monkey made, i.e., the first object he fixated for 0.5 s. To calculate performance, we obtained the percentage of trials in which the monkeys' primary choices were the novel objects.

For single neuron analyses (Figure 2.5B–E) of novelty, task-relevance, and sequence-violations in the object sequence task, we subtracted the activity 100 ms before the object presentation from the activity measured after the object was presented (the time window was 200 ms to 400 ms unless otherwise stated). In this way, changes in firing rate that were unrelated to the objects were not considered in the analyses.

Neuronal discrimination of object novelty was assessed by calculating area under the receiver operating characteristic (ROC) curve. ROC areas of 0 and 1 are equivalent statistically; both indicate that two distributions are completely separated. The analysis was structured so that ROC area values greater than 0.5 indicate that the activity during novel object presentation was greater than familiar.

2.3 Results

2.3.1 CS-related phasic and ramping activity are observed in mostly distinct BF cell groups that differentially signal reinforcement statistics

We recorded BF neurons in 5 monkeys that participated in a Pavlovian procedure in which they experienced reward predictions that varied in magnitude and probability (Monosov and Hikosaka, 2013; Monosov et al., 2015; White and Monosov, 2016). A reward-probability block contained five conditioned stimuli (CSs) associated with five probabilistic reward predictions (0, 25, 50, 75, and 100% of 0.25 mL of juice). A reward-amount block contained five other CSs associated with certain reward predictions of varying reward amounts (0.25, 0.1875, 0.125, 0.065, and 0 mL). During neuronal recording, any neuron that displayed ramping and/or phasic burst responses in the CS epoch of this Pavlovian procedure was recorded (n = 70; monkey H = 15, monkey P = 16, monkey B = 10, monkey R = 12, and monkey Z = 17).

Example neurons are shown in Figure 2.1A. The first neuron (Figure 2.1A, **top**) displayed short latency bursting after the presentation of the probability and amount CSs. This phasic activation was greatest following the presentation of the CS associated with the highest expected value in either the reward-probability or the reward-amount block and least following the presentation of the CSs associated with the lowest expected value (no reward). In either block, the bursting activity was strongly correlated with the expected value (Spearman's rank correlation; probability block, $\rho = 0.84$, $p < 0.0001$; amount block, $\rho = 0.86$, $p < 0.0001$). The second neuron (Figure 2.1A, bottom) had a very different response. Shortly after the CSs were presented, it displayed a consistent CS-onset-related inhibition that was greatest in the low-value trials and less apparent during high-value trials, on average roughly scaling with the expected value. In the reward-

probability block, this initial change was followed by ramping activity to the time of the uncertain (or risky) reward delivery (following 75%, 50%, and 25% CSs). The neuron's activity was significantly fit by a model of uncertainty ($\rho = 0.77$, $p = 0.0001$; measured in the last 500 ms) but not expected value ($\rho = -0.01$, $p = 0.94$). In the reward-amount block, in which all trials were certain, the neuron represented the expected value until the time of the reinforcement in its tonic activity ($\rho = 0.70$, $p < 0.0001$; measured in the last 500 ms). These example neurons suggest that the BF may contain functionally distinct classes of neurons: phasic bursting neurons that co-vary with the magnitude and probability of reinforcements and tonic neurons that ramp, predicting the timing of uncertain outcomes.

To test this, we clustered BF neurons based on their average responses. Only neurons that had been recorded in every condition in both blocks were included (n = 66/70). Importantly, their response vectors were obtained by averaging the neuronal activity across all five CSs in the reward-probability block and were subsequently normalized from 0 to 1 (Figure 2.1B, inset). Therefore, neuronal tuning (e.g., representation of reward probability) and baseline firing rates were not considered in the clustering analysis.

This analysis revealed two clusters (Figure 2.1B). The first cluster (red; n = 23) showed clear bursting after the CS onset (see the neurons' response vectors in Figure 2.1B, inset). In contrast, the second cluster (blue; n = 43) showed an initial suppression following the CS onset and a slow ramp-like increase in activity as the trial's outcome neared.

The two clusters had different baseline firing rates (Figure 2.1C): one had relatively high firing rates (blue cluster; average frequency = 18 Hz; SD = 12 Hz) and the other low (red cluster; average frequency = 2.1 Hz; SD = 3.5 Hz). Both clusters' initial CS responses co-varied with the

magnitude of the predicted reward, initially coding expected value, but the latency of this information was different among the two clusters. The expected value was conveyed earlier by the neurons in the phasic bursting red cluster (Figure 2.1C, right; rank-sum test; $p < 0.01$; blue cluster, average = 195 ms, median = 159 ms, SD = 104 ms; red cluster, average = 123 ms, median = 100 ms, SD = 96 ms).

These clusters differed in how they represented both the probability and amount of reinforcement (Figures 2.1D–2.1F and Supplemental Figure 2.1). Phasic bursting neurons (red cluster) signaled the expected value of the CSs in their bursting activations. The bursting activity was correlated with the probability in the probability block ($\rho = 0.60$, $p < 0.0001$) and with the reward amount in the amount block ($\rho = 0.71$, $p < 0.0001$). Tonic ramping neurons' initial suppression co-varied with the expected value ($\rho = 0.47$, $p < 0.0001$ in the probability block; $\rho = 0.46$, $p < 0.0001$ in the amount block). However, in trials in which reward was uncertain, they displayed additional ramping activity toward the trial outcome (Monosov et al., 2015). The activity during these 75%, 50%, and 25% CS trials was correlated with the probability of reinforcement delivery (Spearman's rank correlation; $\rho = 0.25$, $p = 0.0043$; pre-outcome analysis window −0.5 s before the outcome was delivered). And, on average, the pre-outcome activity in the reward-probability block (across all 5 trial types) was correlated with uncertainty ($\rho = 0.71$, $p < 0.0001$; same analysis window as above).

Locations of phasic bursting and tonic ramping neurons were reconstructed using in vivo MRI (Materials and Methods) (Daye et al., 2013) (Supplemental Figure 2.2). Both phasic bursting and tonic ramping neurons were found within the BF, in the diagonal band of Broca and the nucleus basalis of Meynert (Mesulam et al., 1983; Monosov et al., 2015; Turchi et al., 2018).

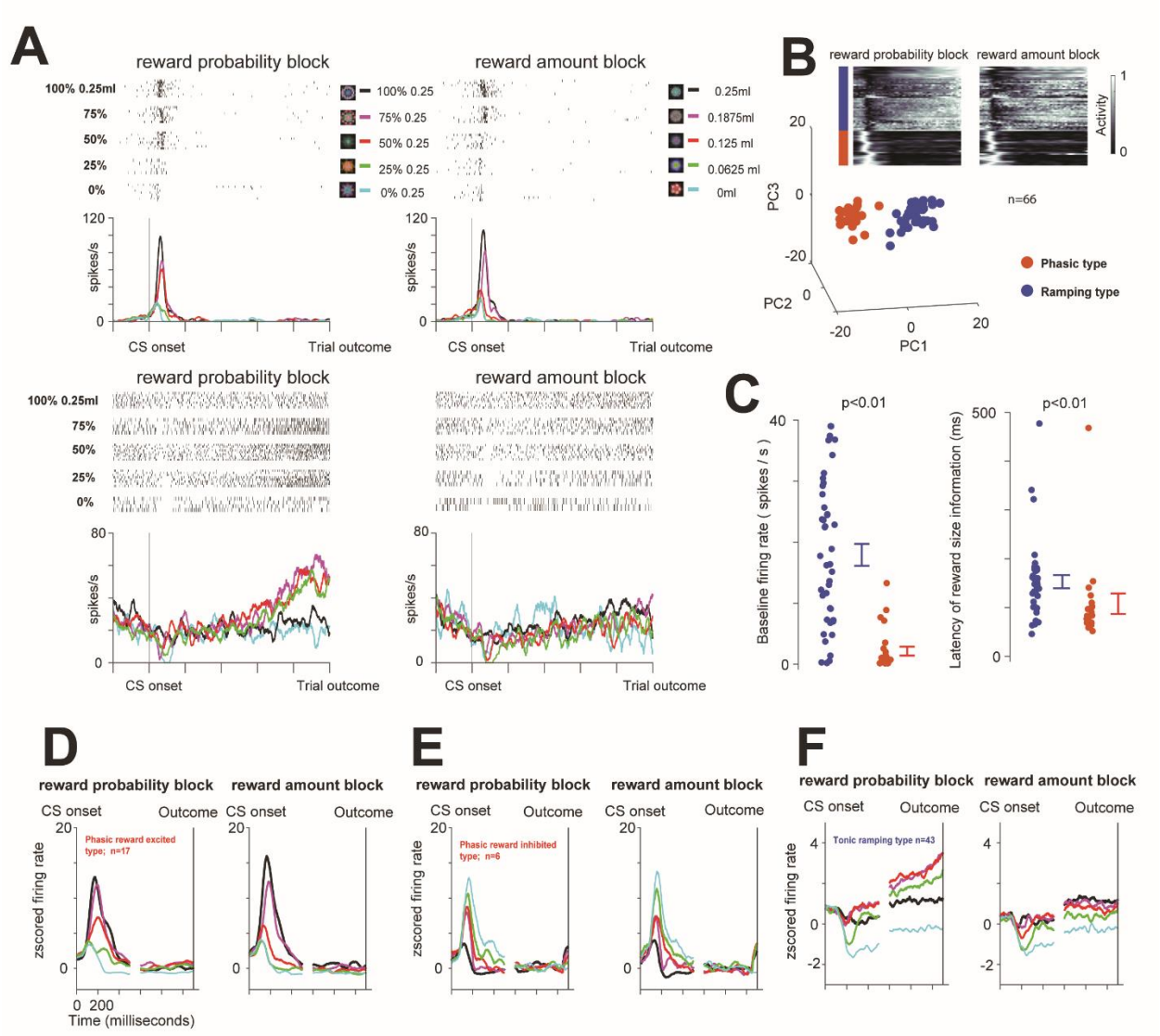


Figure 2.1. Two groups of BF neurons encode the magnitude and probability of reinforcement in distinct manners. (**A**) Responses of two example BF neurons (top and bottom) to the presentation of 10 fractal objects associated with certain and uncertain predictions of juice rewards in the reward-probability block (**left**) and reward-amount block (**right**). (**B**) Clustering of BF neurons based on average activity in the probability block. The inset heatmap shows the activity of 66 BF neurons (normalized from 0 to 1 to the minimum and maximum in the reward-probability block) from the time of the CS onset to the time of the trial outcome (reward or no reward) in the reward-probability and reward-amount blocks. Each line represents the average activity across all 5 trial types in the block for each neuron. Below are the results of principal-component analyses performed on those normalized CS response functions. K-means clustering (Materials and Methods) was used to separate the neurons into two groups: red group (n = 23) and blue group (n = 43). The group identities of the neurons are also indicated by a color bar to the left of the heatmap. (**C**) The two clusters of neurons (red and blue) display distinct baseline firing rates (**left**) and latencies of value coding (**right**) in the reward-amount block. Each

dot represents data from a single neuron. Error bars around the mean show the SEM. (**D-E**) Average responses of the neurons in the red group in the reward-probability block (**left**) and reward-amount block (**right**). (**D**) shows neurons that displayed greater activation for reward versus no-reward trials, while (**E**) shows neurons that displayed greater activation for no-reward trials. See also Supplemental Figure 2.1, Materials and Methods, and the associated Supplemental Figure 2.2 for details and anatomical locations of neuronal recordings. (**F**) Average responses of the neurons in the blue group in the reward-probability block (**left**) and reward-amount block (**right**).

2.3.2 Phasic and ramping neurons signal early versus late rewards under temporal uncertainty

Does ramping of BF neurons encode the estimated timing of uncertain rewards? If so, then if rewards were certain but their timing was uncertain, the neurons should display ramping activity to the time of the earliest possible reward. Second, phasic bursting neurons' bursts seemed to scale with the expected values of the CSs, regardless of whether the value was manipulated by probability or amount. Might these neurons also encode the value of early versus late rewards?

To answer these questions, we designed a reward-timing procedure (Supplemental Figure 2.3). Here, five distinct visual-fractal objects served as CSs that predicted either (1) a probabilistic delay before a reward with deterministic delivery (delays = 1.5 or 4.5 s; reward-timing-uncertain CSs) or (2) a deterministic delay before a reward with 0.5 probability of delivery (reward-probability CS). To test how phasic bursting neurons and tonic ramping neurons encode temporally uncertain reward predictions, we first identified them using the task in Figure 2.1 and then recorded them in this reward-timing procedure (n = 52; monkey W = 21, monkey B = 6, monkey R = 15, and monkey Z = 10).

Tonic ramping neurons displayed ramping activity in the 0.75, 0.5, and 0.25 reward-timing-uncertain conditions (Supplemental Figure 2.3). The magnitude of this activation was correlated with the probability of reinforcement delivery at 1.5 s (Spearman's rank correlation; $\rho = 0.48$, $p < 0.0001$; analysis window, 1 s to 1.5 s). Interestingly, significant ramping was also observed to certain late reward at 4.5 s (Supplemental Figure 2.3). Therefore, BF ramping tracks reward delivery during temporal-reward uncertainty (before 1.5 s) and during relatively longer epochs in which there is temporal uncertainty due to noise in interval timing.

Phasic bursting neurons' activity scaled with reward timing such that highest activity was evoked by CSs predicting the earliest reward (Supplemental Figure 2.3). Their average activity was correlated with reward probability at 1.5 s (Spearman's rank correlation; $\rho = 0.41$, $p = 0.0012$; analysis window, 0 s to 0.5 s). Unlike the tonic ramping neurons, the phasic bursting neurons did not anticipate the late reward at 4.5 s (Supplemental Figure 2.3).

2.3.3 BF phasic and ramping neurons signal reinforcement surprise in distinct manners

A long-standing question is whether the BF signals errors in state values, or RPEs—a key signal for updating reward values and mediating economic choice (Schultz, 2002; Lak et al., 2014). An alternative is that BF neurons signal a rectified (unsigned) prediction error (Pearce and Hall, 1980; Roesch et al., 2010) rather than a value (signed) prediction error, which is better suited to control attention and mediate memory of salient events. We tested which type of prediction error is signaled by the BF by analyzing responses to reward deliveries and reward omissions after 25%, 50%, and 75% predictions (Figure 2.2).

Ramping neurons' outcome-related activity on average was correlated with unsigned prediction errors (Figures 2.2A and 2.2B). After the trial outcome, the magnitude of their activity was greatest during reward-delivered trials following 25% reward predictions and greatest during reward-omission trials following 75% reward predictions. Reward-omission and reward-delivery outcome responses were significantly correlated with expectancy (Figure 2.2B), albeit in opposite manners.

Phasic bursting neurons' outcome-related activity also signaled prediction errors following reward deliveries. Their delivery responses were correlated with expectancy (Figure 2.2D, red),

displaying highest activations following reward deliveries in 25% reward trials. However, unlike the ramping neurons, these neurons did not discriminate reward omissions following different uncertain reward predictions (Figures 2.2C and 2.2D). To verify that the lack of relationship between reward-omission-related activity and reward probability was not due to firing rate normalization, we repeated the correlation analyses in Figure 2.2D on raw omission-related spike counts and observed the same results ($p = 0.89$). Hence, a key feature of the value RPE—a reward-omission-related suppression—was missing from the phasic bursting neurons.

BF bursting can be elicited by rewarding and aversive, noxious events(Lin and Nicolelis, 2008; Hangya et al., 2015; Monosov et al., 2015). Therefore, why was bursting not apparent in response to unexpected reward omissions? The most parsimonious explanation is that the lack of omission responses was due to a lack of external salient events cueing reward omissions and a lack of sensitivity of phasic bursting neurons to internally generated errors in subjective value.

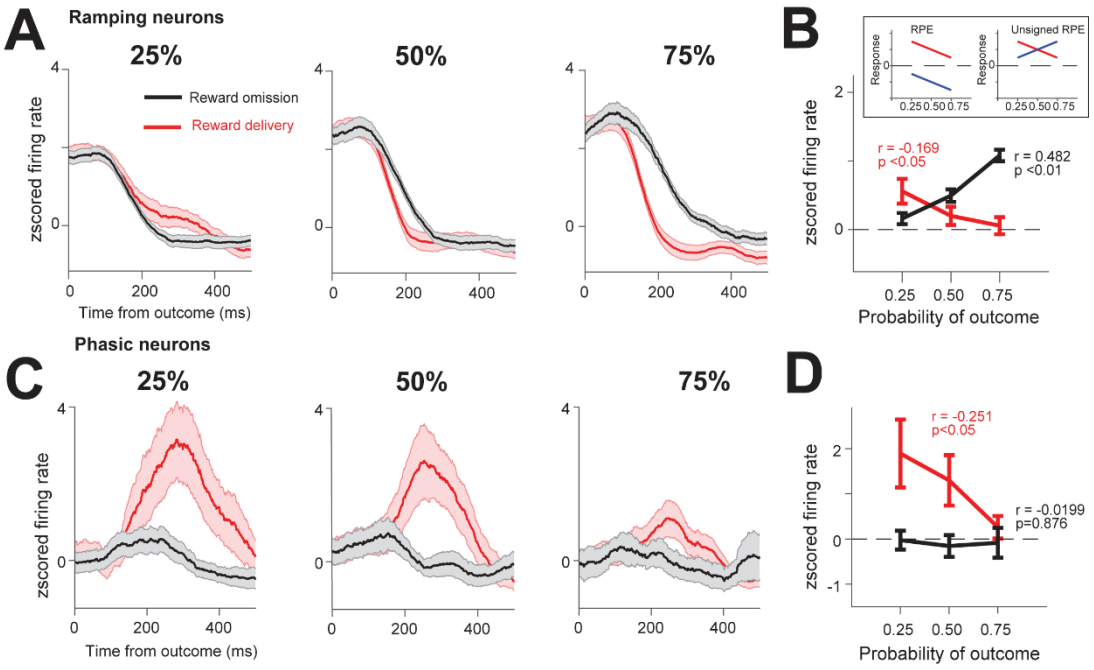


Figure 2.2. Differential coding of surprise in BF ramping and bursting neurons. (**A**) Ramping neurons' average outcome activity in 25%, 50%, and 75% conditions. Red shows reward-delivered trials; black shows no-reward trials. (**B**) Ramping neurons' average responses for reward-delivery and no-reward trials. Linear correlations of responses with reward expectancy are indicated (time window: 100 ms to 400 ms; p values were obtained with 10,000 permutations; Materials and Methods). The results of the correlations suggest that the activity resembles the toy model of unsigned RPEs (or surprise). The inset shows cartoon models of theoretical outcome responses coding RPEs (**left**) and unsigned RPEs (**right**). If neurons signal unsigned RPEs, then they should display greatest responses to reward deliveries following 25% reward predictions and smallest responses following a 75% reward prediction. The same neurons should display greatest responses to reward omissions following 75% reward predictions and smallest responses following 25% reward predictions. Alternatively, if neurons encode signed RPEs, then they will display inhibitions following omissions whose magnitudes ought to be inversely related to the probability of a reward. (**C**) Outcome activity of phasic bursting neurons. Conventions are the same as in (A). (**D**) Phasic bursting neurons' responses resembled RPE coding only in reward-delivery trials (red; time window: 200 ms to 500 ms).

Surprise has a temporal dimension, and ramping neurons clearly display ramping signals to the timing of uncertain or salient reinforcements (Figures 2.1 and 2.2) (Monosov et al., 2015), the magnitude of which is correlated with monkeys' confidence in reward delivery (Figures 2.1 and Supplemental Figure 2.3; ramping responses: $0.25 < 0.5 < 0.75$). Might BF ramping activity encode estimates of outcome timing under uncertainty?

To test this, we took advantage of the fact that in our tasks, CSs co-terminated with outcomes. During omission trials, no external cues indicated that the reward was omitted. If ramping reflects information about the animals' internal temporal estimates, then we should have seen different ramping-down responses following omissions in 25%, 50%, and 75% trials.

BF ramping returned to baseline earliest during 25% reward trials and latest during 75% trials (Figure 2.3A). Decay of the ramping also roughly scaled with reward expectation: it was greatest following omissions during 25% and least during 75% trials (Figure 2.3B; bootstrapping; the 95% confidence intervals of 25%, 50%, and 75% decay rates exclude each other). Note that different firing rates across different trial types could not explain these results because before obtaining the decay rates, we first normalized each trial type from 0 to 1.

Next, we studied the activity of BF ramping neurons in the reward-timing procedure because it contained two distinct 50% reward predictions: one in which the CS co-terminated with the outcome and one in which the CS remained on the screen (Supplemental Figure 2.3). In the first condition, the animals obtained a signal about the timing of the trial, while in the other, they did not. The decay rate of BF ramping neurons was again sensitive to temporal predictions: it was greater when the animals did not receive an explicit temporal cue (Figure 2.3C; bootstrapping; the 95% confidence intervals of the decay rates exclude each other). Finally, we analyzed

another task that contained two types of 50% reward CS trials with identical timing and reward statistics. The two trials differed in one way—one of them contained an external trial-end cue that indicated when the trial was over. Consistent with the results of Figure 2.3C, when an explicit cue was given during reward-omission trials, the ramping-down activity displayed a relatively rapid drop-off (Figure 2.3D, right). In sum, Figure 2.3 show that BF ramping activity is strongly influenced by evidence about and confidence in the timing of reinforcements.

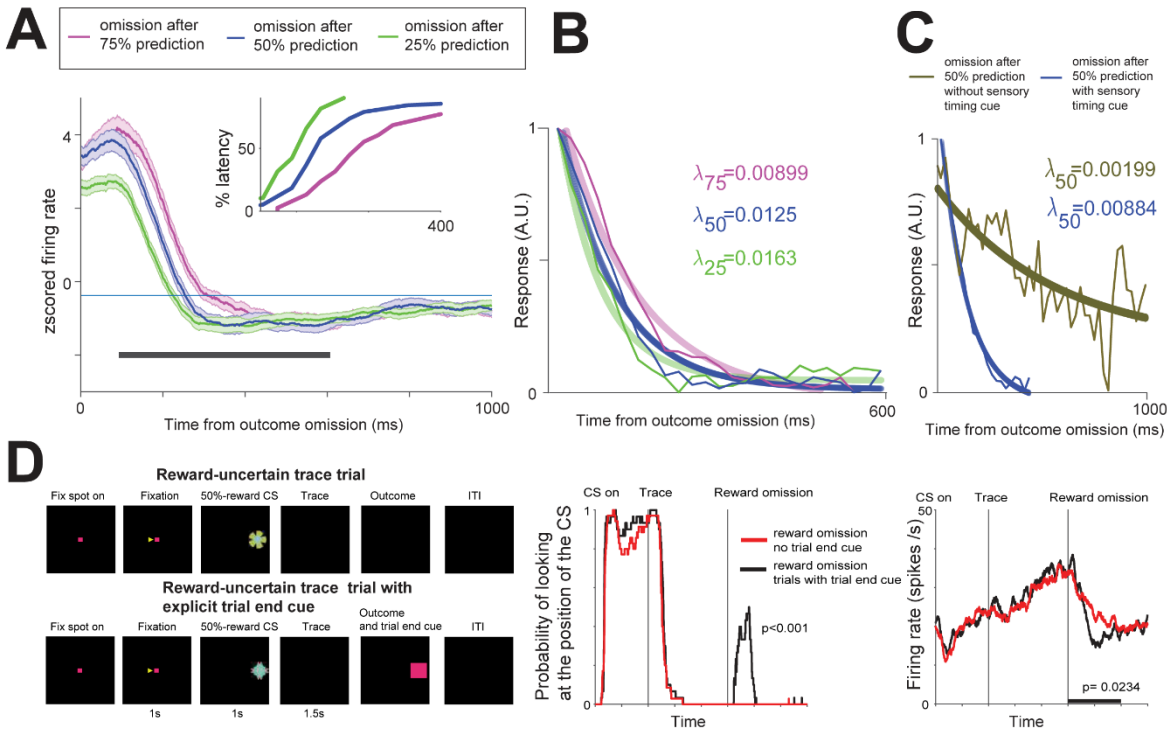


Figure 2.3. BF ramping neurons encode estimates of outcome timing under uncertainty. (**A**) Activity of BF ramping neurons during 25%, 50%, and 75% reward-probability trials in which the reward was omitted. The ramping activity returned to the inter-trial baseline level (thin blue line) at different latencies across these three types of trials: earliest during 25% trials and latest during 75% trials. Cumulative distributions of these latencies are shown in the inset. The black bar below the activity indicates the time window for the analyses in (B). (**B**) Exponential fits (thick lines) to the population's binned activity (thin lines; Materials and Methods). Fits and decay rates (**right**) were calculated for the population after the activity for each trial type was normalized from 0 to 1, such that for each of the three conditions, the starting point is 1. A.U., arbitrary units. (**C**) Same as (B), except here we compared the fit and decay rate during 50% trials in which an explicit cue indicated the end of the trial (dark blue) with the fit and decay rate during 50% trials in which no explicit cue was given (and the CS remained on the screen; Materials and Methods). (**D**) **Left**: trace conditioning with and without explicit visual cues that signaled the end of the trial. Middle: the monkey's gaze behavior indicated that it attended to the trial-end cue (presented at the same location as the CS; rank-sum test; $p < 0.001$). **Right**: explicit knowledge of trial timing reduced the reward-omission-related ramping activity (monkey W; 8 neurons; $p = 0.0234$; signed-rank test). The analysis window used to study gaze behavior and neuronal activity is indicated by the black bar. The shaded regions throughout this figure represent the SEM. See also Supplemental Figure 2.3 for activity in the temporal-uncertainty procedure separately.

2.3.4 Object novelty and sensory surprise are signaled by the BF

The data thus far show that BF neurons are sensitive to surprise. However, surprises arise due to violations in belief states following a probabilistic prediction, when there is a deviation of the outcome from the mean of expected-outcomes (Barto et al., 2013), or as a result of novelty due to a comparison of a sensory events with representations of past experiences. To test how the BF represents novelty, we designed an object-sequence task in which novel objects were fully expected.

Monkeys experienced four sequences of object presentations (S1, S2, S3, and S4). Each sequence contained 3 familiar objects and 1 novel object. The novel object was always in the second position in the sequence. If a neuron has a selective novelty response, it should respond more strongly and consistently to the novel object than to the familiar objects in the sequence. To assess whether novelty responses were dominantly due to task relevance or reward prediction, following S2 and S4, monkeys performed a reaction-time delayed non-matching-to-sample (DNMS) task (Figure 2.4A, right). During the DNMS task, an object that was novel during the presentation of S2 (or S4 if the DNMS trial followed S4) was presented along with a novel object that had never been experienced. The trial continued until the monkeys fixated on this novel object for 0.5 s to get a reward (Figure 2.4A, right). The monkeys' behaviors indicated that they understood the task and utilized previous experiences to increase their reward rate. Their first saccade following the presentation of the two fractals most often landed on the novel object, where their gazes remained until the non-selected stimulus disappeared and a reward was delivered.

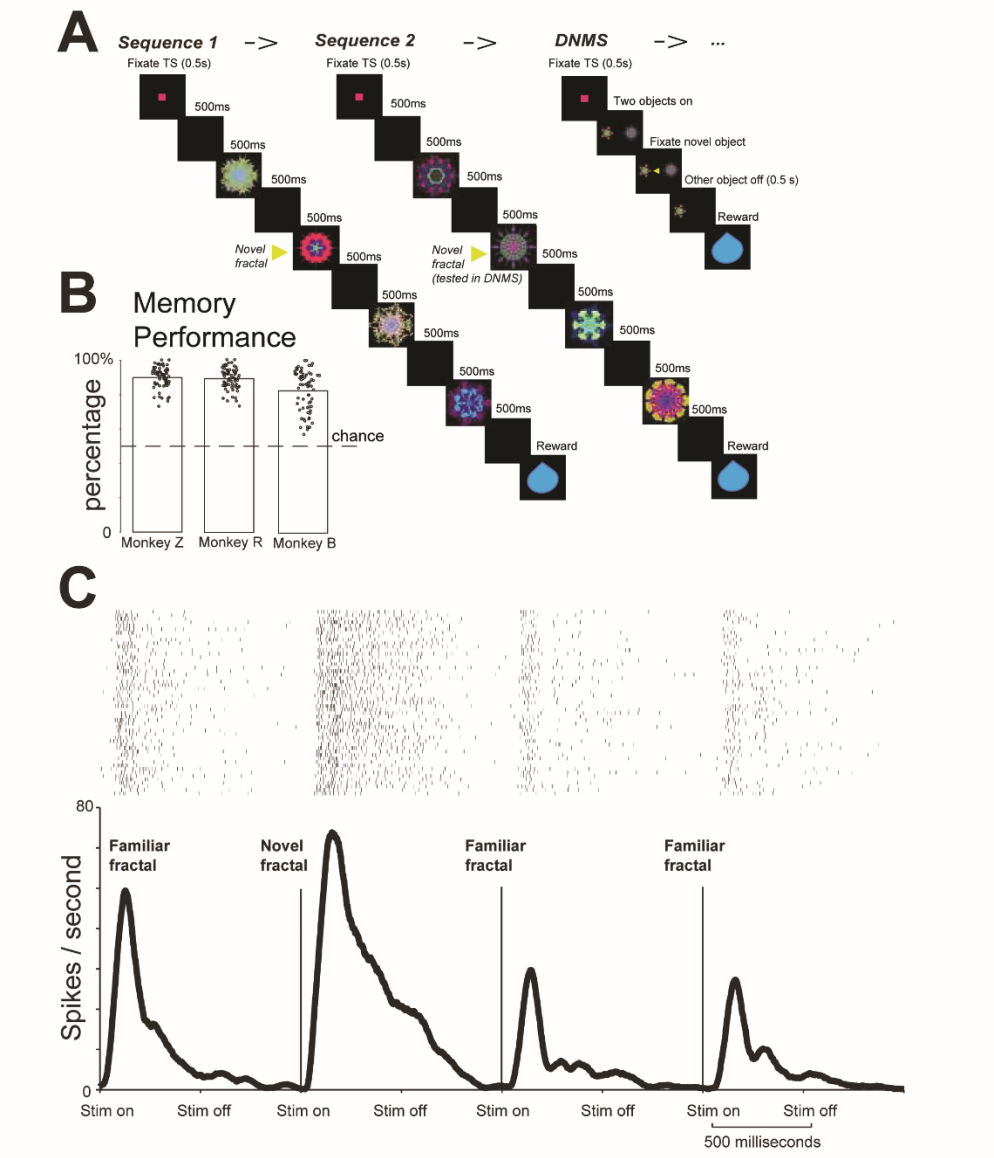


Figure 2.4. Object-Sequence Task. (**A**) The monkey was first shown sequences of fractals. Each sequence contained 4 fractals, in which the 1st, 3rd, and 4th fractals were fixed familiar objects and the 2nd fractal was always novel. After the two sequences, the monkeys performed a DNMS task in which one object was novel and the other was the object that was previously novel in sequence 2. Monkeys fixated the novel object for reward. (**B**) Behavioral performance for three monkeys. y axis shows the percentage of first saccades to the novel object in DNMS. The percentages are significantly different from 0.5 for all three monkeys ($p < 0.01$; signed-rank test). (**C**) Example BF phasic bursting neuron's responses to the four objects in a sequence. The response was highest for the second (novel) fractal (rank-sum test; $p < 0.05$).

We studied 39 BF neurons identified using experiment 1 (monkey B = 6, monkey R = 11, and monkey Z = 22). Phasic bursting neurons robustly discriminated the novel object from the familiar objects. An example phasic bursting neuron is shown in Figure 2.4C. This neuron responded selectively to the novel object ($p < 0.01$; rank-sum test). This selective response could not be explained by priming or reward proximity because the novel objects always appeared in the second position in the sequence (Figure 2.4A) rather than the first or the last. Like the example neuron, the population of phasic bursting neurons (Figure 2.5A) and the single neurons (Figure 2.5B) selectively discriminated the novel object versus familiar objects.

Phasic bursting neurons' strong and selective novelty responses in the object-sequence task were present when the novel object was relevant or irrelevant for subsequent memory behaviors (Figure 2.5C). That is, during both S1 and S3, BF phasic neurons displayed stronger responses to novel objects than to familiar objects (signed-rank tests; $p < 0.01$). Their novelty responses were also consistently enhanced by task relevance (Figures 2.5C and 2.5D).

An important consideration for the interpretation of novelty responses is that novelty, in primates, is thought to exert a strong influence on behavior (Berlyne, 1970; Tiitinen et al., 1994; Barto et al., 2013; Foley et al., 2014), especially on gaze behavior. However, the type of influence (attentional, motivational, or both) that is exerted has been unclear. We designed a novel behavioral procedure that revealed that object novelty indeed has a motivational value (Supplemental Figure 2.5). This finding necessitates that future studies assess the role of BF activity in mediating the motivational effects of object novelty on behavior.

We previously showed that CSs predicting uncertain (surprising) rewards attract overt attention more than CSs predicting certain rewards (Monosov, 2017). Here and in a previous report

(Monosov et al., 2015), we showed that BF ramping neurons anticipate uncertain reward delivery (Supplemental Figures 2.1 and 2.3). So, might these neurons also anticipate other attention-capturing stimuli such as novel objects? While in contrast to the phasic neurons, the ramping neurons had a weaker novelty-selective response (rank-sum test comparing single neurons' area under ROC curve values; $p = 0.035$), they indeed displayed ramping that anticipated at least two critical events in the object sequence task: the presentation of novel objects and rewards occurring after a long interval (Supplemental Figure 2.4).

The temporal cortex, a major target of BF projections (Mesulam et al., 1983), is sensitive to sequence violations (Meyer and Olson, 2011). To test whether the BF is sensitive to unexpected violations in object sequences, we replaced an object in S2 with an object from S1 or an object from S4 with an object from S3 in ~11% of trials. These replacements avoided RPEs because the proximity to the reward was not changed. Sequence violations produced small but significant increases in the population responses of phasic and tonic neurons (Figure 2.5 and Supplemental Figure 2.4). So, the BF can broadcast information about novel and surprising sensory events that are not directly associated with primary reinforcements.

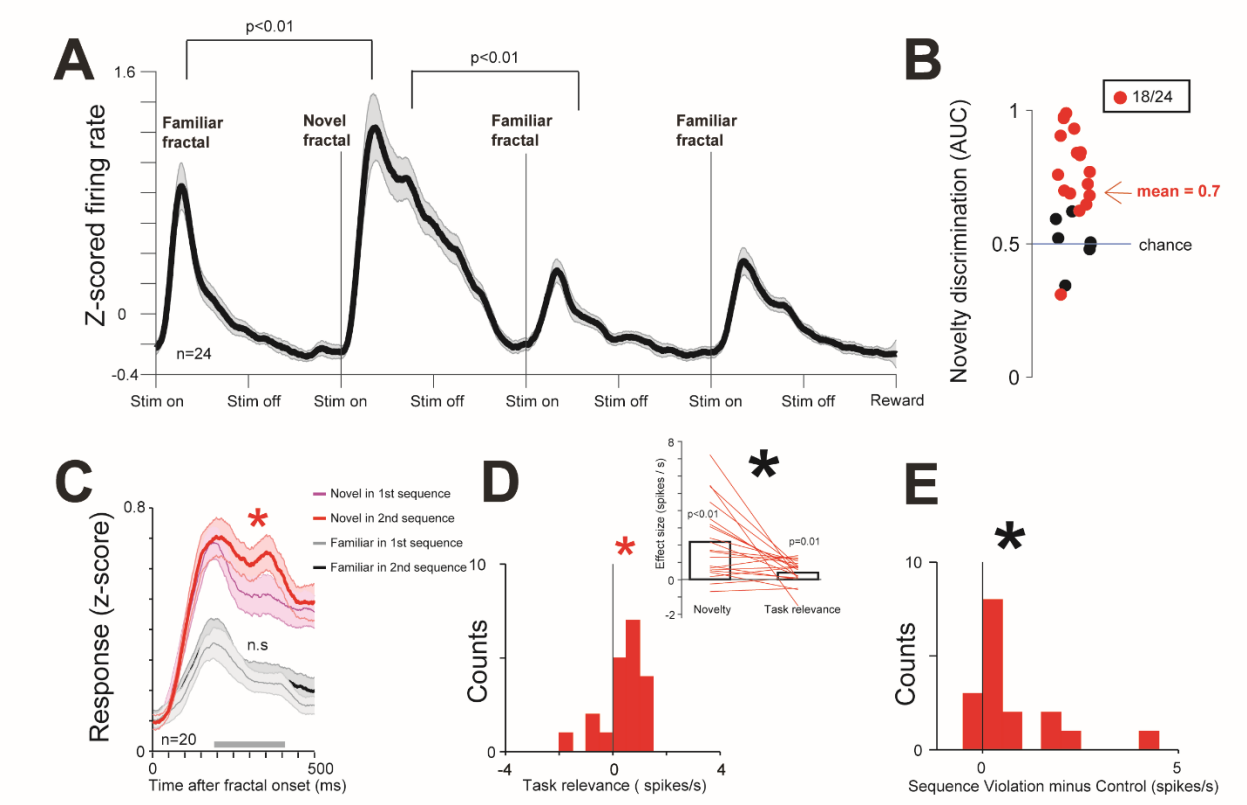


Figure 2.5. Phasic Bursting Neurons Signal Novelty and Surprise Not Directly Related to Reward. (**A**) Average activity of phasic bursting neurons in the object-sequence task. The shaded region represents the SEM. (**B**) Area under the ROC curve (AUC) for each phasic neuron that assessed the ability of the neuron to discriminate novel versus familiar objects. Red dots are neurons that can significantly discriminate novel versus familiar objects (time window: 200 ms to 400 ms). (**C**) Phasic neurons group average responses to novel fractals in sequence 1 (thin blue line), sequence 2 (thick red line), and then to the last 2 familiar fractals in sequence 1 (thin gray line) and sequence 2 (thick black line). The shaded region represents the SEM. The asterisk indicates significant difference ($p < 0.05$) between novel fractal responses in sequences 1 and 2. n.s., not significant. (**D**) Lower left: histogram of single neurons' response differences for novel fractals in sequence 2 (or 4) and sequence 1 (or 3). The red asterisk indicates a significant difference from 0 ($p < 0.05$). **Upper right**: for each neuron, the data from the histogram (right) was compared with the strength of novelty discrimination (left). Both novelty-discrimination and task-relevance effects are significant, but the novelty effect is stronger ($p < 0.05$). (**E**) At low probability (11%), one of the familiar fractals in sequence 2 (or 4) was substituted with another familiar fractal from sequence 1 (or 3) (Materials and Methods). Phasic neurons' responses were enhanced ($p < 0.01$) by this object-sequence violation. See also Supplemental Figure 2.4 for the activity of tonic ramping neurons.

2.4 Discussions

We report that the primate BF contains at least two types of neurons that process a diverse set of salient events in distinct manners: with phasic burst responses when they occur or with ramping activity—in anticipation of their occurrence.

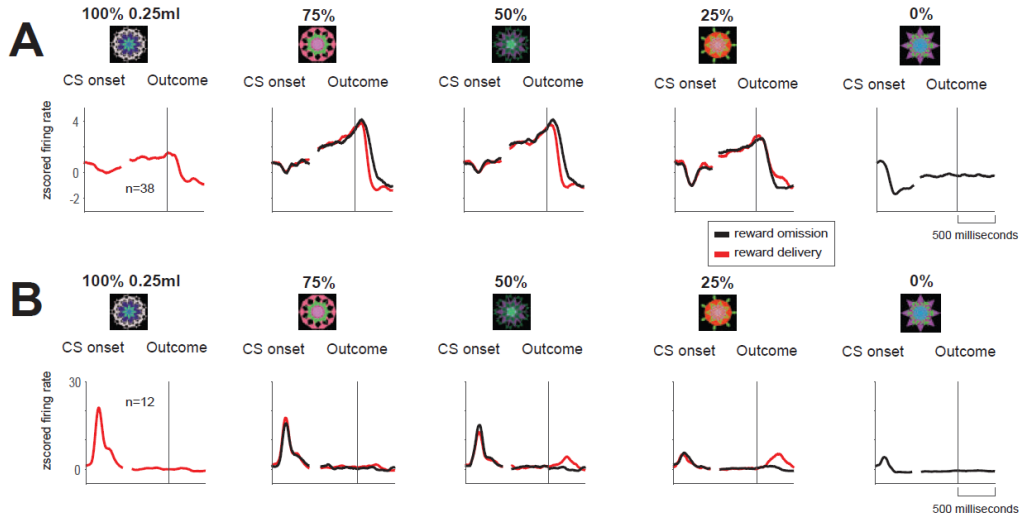
Ramping neurons signaled internal variables closely tied to confidence in the timing of surprises and novel events. Their activity may represent (or provide a readout of (Paton and Buonomano, 2018)) an internal clock that is well-suited to guide anticipatory temporal attention, particularly in uncertain or novel contexts. Phasic bursting neurons rapidly and precisely conveyed statistical information about the timing, magnitude, and probability of reinforcement predictions and about the surprise of reinforcement deliveries. They were highly sensitive to sensory novelty and to errors in the subjects' beliefs about the sequences of sensory events. These neurons' short latency bursting could rapidly coordinate many regions of the neocortex that receive BF projections to mediate the processing of a wide range of external salient events and orchestrate appropriate responses to them (Shuler and Bear, 2006; Hangya et al., 2015; Raver and Lin, 2015; Liu et al., 2017; Paton and Buonomano, 2018; Turchi et al., 2018).

Phasically bursting neurons did not discriminate among expected and unexpected reinforcement omissions that monkeys had to detect internally (e.g., omissions were not cued). Thus, in contrast to many dopamine neurons, they did not convey phasic RPEs (Morris et al., 2004; Matsumoto and Hikosaka, 2009; Lak et al., 2014). Notably, a set of recent studies showed that not all dopamine-phasic responses signal RPEs wholly or purely. Instead, some dopamine neurons convey an alerting signal complementary to BF bursting (Bromberg-Martin et al., 2010a; Matsumoto and Takada, 2013; Takahashi et al., 2016; Takahashi et al., 2017; Babayan et al.,

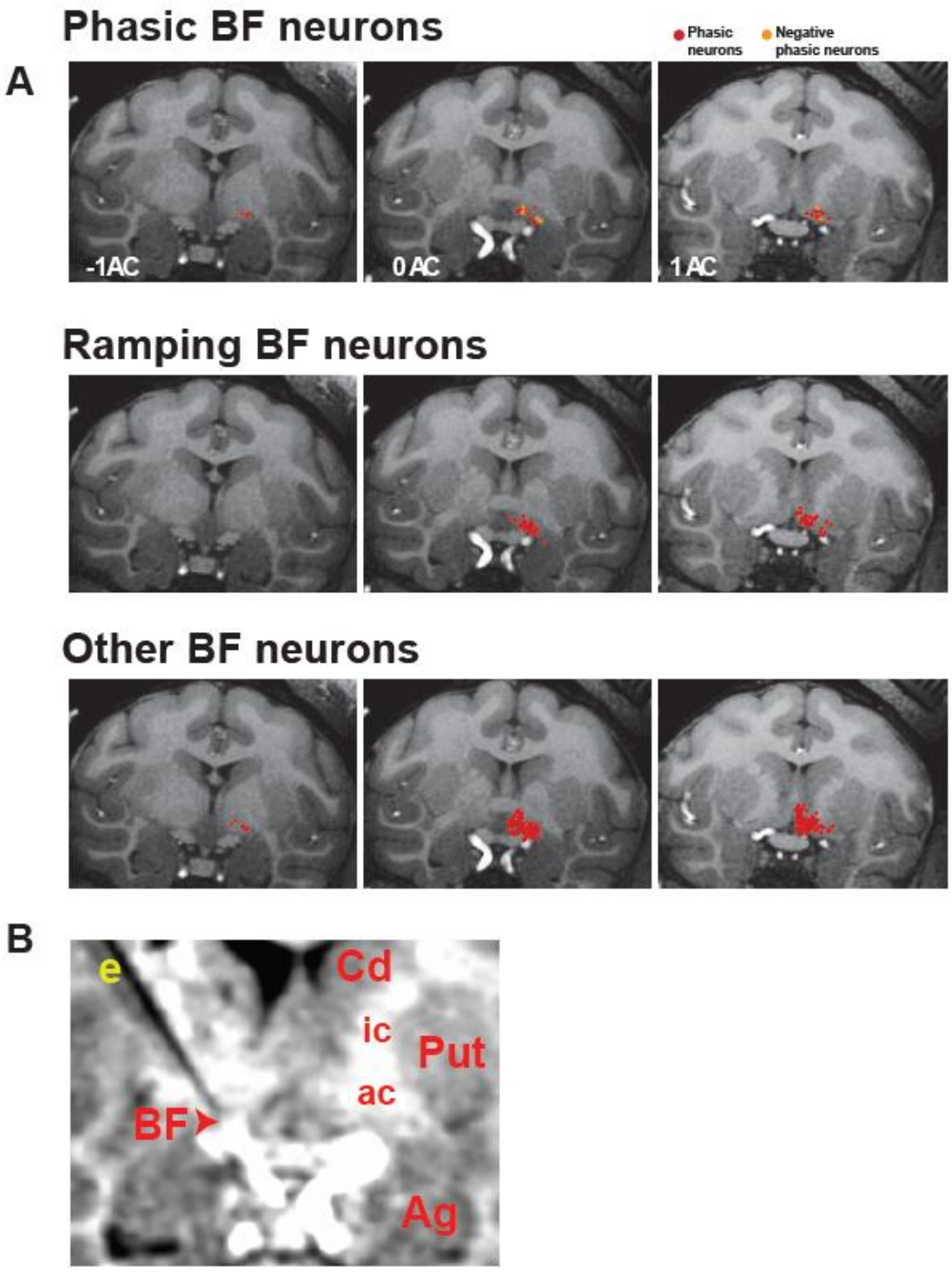
2018). Future studies must assess how the BF phasic bursting and dopamine neurons work together to mediate behavior. One possibility is that BF phasic bursting (conveyed to the neocortex in response to a salient event) is followed by the release of dopamine in the basal ganglia. This dopaminergic release would then support striatal value (or motivational-salience) assignments to events being processed by the cortex (under the mediation of the BF). How dopamine would do so may ultimately depend on when and where it is released (Bromberg-Martin et al., 2010a; Matsumoto and Takada, 2013; Takahashi et al., 2016; Takahashi et al., 2017; Babayan et al., 2018).

The BF contains prominent groups of cholinergic, GABAergic, and glutamatergic projection neurons. Previous work in rodents has identified putative GABAergic CS-related phasic bursting neurons, reinforcement-salience-related bursting cholinergic neurons, and other tonically active neurons in the rodent BF (Lin and Nicolelis, 2008; Avila and Lin, 2014; Hangya et al., 2015). It will now be particularly important to identify which neurotransmitters are released (or co-released) by phasic bursting and ramping neurons in primates.

2.5 Supplemental Materials

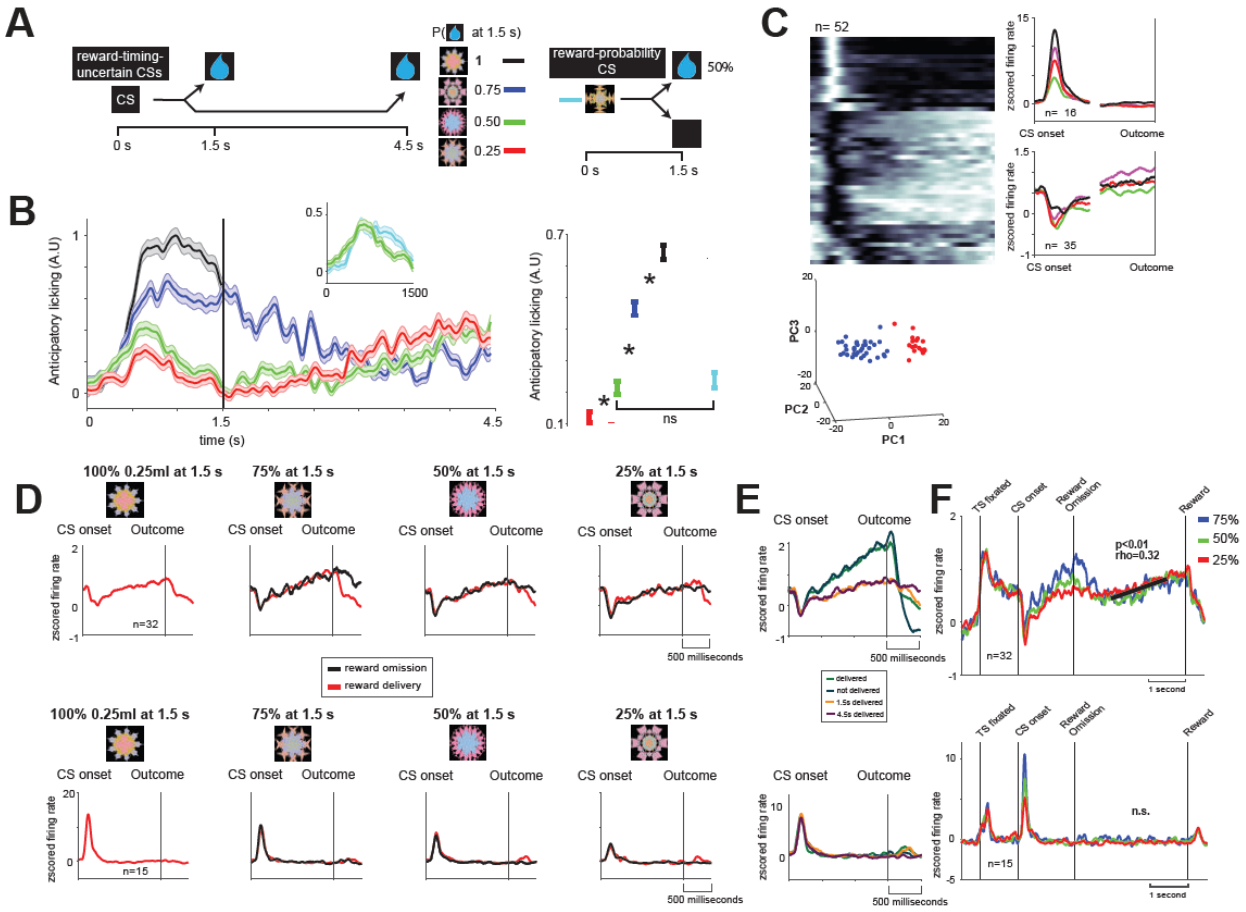


Supplemental Figure 2.1. BF activity across different probabilistic reward predictions. (**A-B**) Neurons' average activity shown separately in trials in which rewards were predicted with 5 different reward probabilities (indicated on the **top**; actual fractals used in the task are shown above the neuronal activity). After the trials' outcome time, activity is shown separately for reward delivered trials (red) and reward omitted trials (black). (**A**) Ramping neurons (**B**) Phasic bursting neurons. In this figure neurons with at least 2 trials for each condition (e.g. delivery versus omission) are shown.



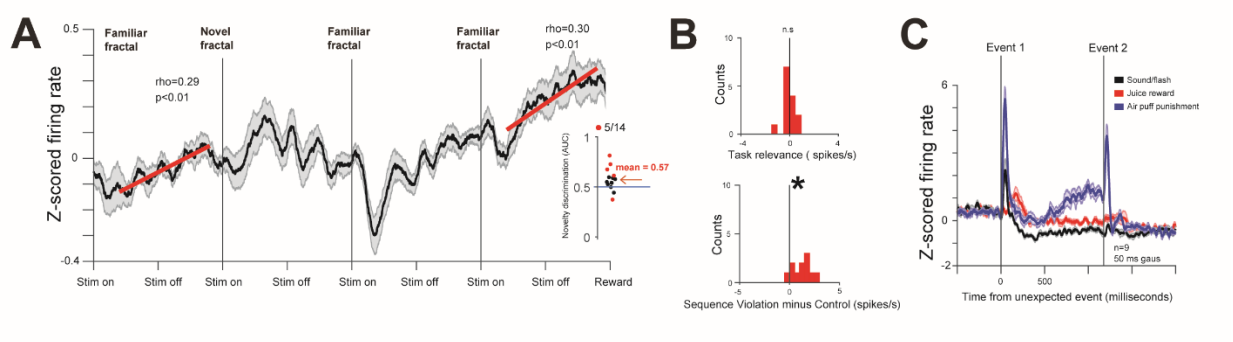
Supplemental Figure 2.2. Estimated locations of phasic bursting and tonic ramping neurons in the BF. (A) The recording range in the BF was -2 to 4 mm anterior to the center of the anterior commissure (AC). Phasic neurons (**top**; n=38), tonic ramping neurons (**middle**; n=79), and other neurons encountered in the BF that did not have ramping or phasic activity

(**bottom**; n=280) are shown on three coronal T1 MRI images. All single neurons across all tasks are shown here. (B) A coronal MRI confirming a recording location of a phasic bursting neuron within the BF of monkey B. The image was acquired with a tungsten electrode (FHC) at the recording location within BF. The electrode's shadow is the black line whose tip is in BF (marked by a yellow e). BF - basal forebrain; ac - anterior commissure; ic - internal capsule; Cd - caudate; Put - putamen, Ag - amygdala.

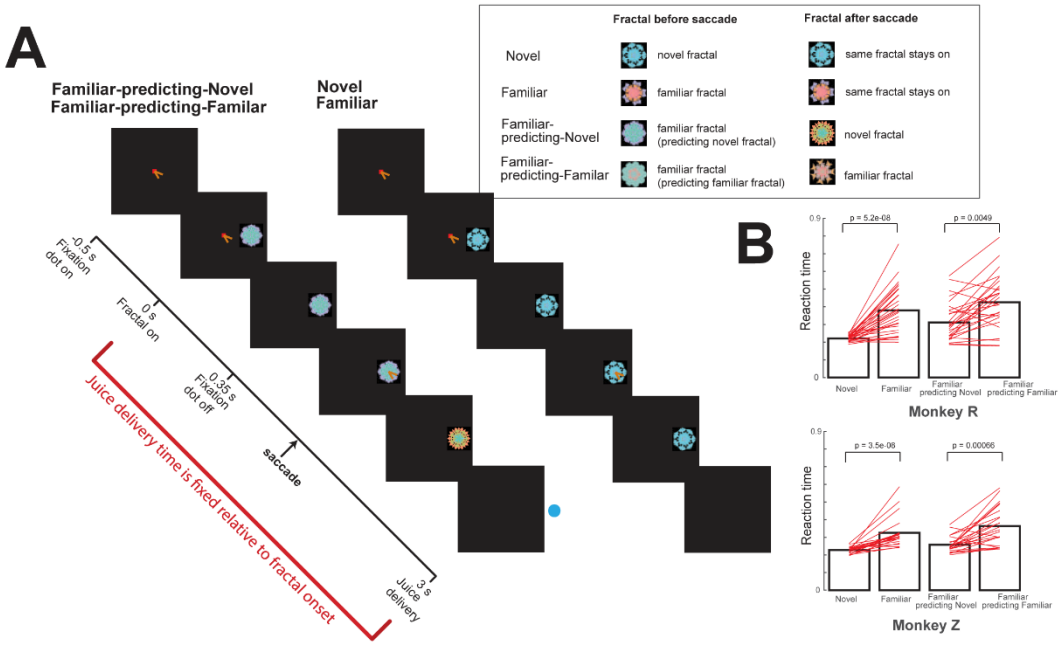


Supplemental Figure 2.3. Anticipatory licking and neuronal activity in reward timing uncertain task. (**A**) The reward timing procedure uses Pavlovian delay conditioning in which five distinct visual fractal objects serve as conditioned stimuli (CSs) that predict either (1) a probabilistic delay before a reward with deterministic delivery (reward-timing-uncertain CSs); or (2) a deterministic delay before a reward with some probability of delivery (reward-probability CS). In trials with one of the four reward-timing-uncertain CSs, reward is delivered at the latest with a delay of 4.5 s after CS onset. However, depending on the CS, reward has either a 0.25 (red), 0.50 (green), 0.75 (blue), or 1 (black) probability of being delivered earlier with a delay of 1.5 s after CS onset. In reward-probability trials, reward is delivered with a delay of 1.5 s after CS onset with 0.50 (cyan) probability. (**B**) Time course of anticipatory licking behavior is shown before possible reward delivery at 1.5 s across all trials; and from 1.5 s to 4.5 s across trials with a reward-timing-uncertain CS in which reward was not delivered at 1.5 s. (**B-right**) The mean magnitude of anticipatory licking behavior increases with the probability of reward delivery at 1.5 s (Spearman's rank correlation, rho=0.38, $p<0.0001$). The asterisks indicate significant differences between CSs (Wilcoxon rank-sum test, $p<0.05$). The 'ns' indicates no significant difference between CSs (Wilcoxon rank-sum test, $p>0.05$). (**C**) Clustering of BF neurons based on average activity in the reward timing task produces similar results to Figure 2.1. Heat map shows the activity of 52 BF neurons (normalized from 0 to 1) from the time of the CS onset to the time of the first trial outcome (1.5 seconds). Each line represents the average activity across all trial types for each neuron. K-means clustering (Materials and Methods) was used to separate

the neurons into two groups: red group (n=17) and blue group (n=35). The first 3 PCSs considered in the clustering are shown at the bottom. The average firing rates of the two groups during the first and last 500 milliseconds of the CS epoch are shown on the right (same format as in Figure 2.1). Note that in the red group (comprising of phasic bursting neurons) there was also 1 negative value neuron whose activity is not shown in the average. (**D**) Tonic ramping neurons' (**top**) and phasic bursting neurons' (**bottom**) average activity shown separately in reward timing uncertain trials in which rewards were predicted at 1.5 s with 4 different probabilities (indicated on the **top**; actual fractals used in the task are shown above the neuronal activity). After the trials' outcome time, activity is shown separately for reward delivered trials (red) and reward omitted trials (black). (**E**) The CS related activity of tonic ramping (**top**) and phasic (**bottom**) BF neurons during 50% reward trials in which reward was delivered or omitted at 1.5 s and 50% reward trials in which reward was delivered at 1.5 s (50% of the trials) or 4.5 s (if it was omitted at 1.5 s). Phasic bursting neurons did not discriminate between these two types of trials ($p = 0.51$, sign rank test, time window: 100ms to 600ms after fractal onset). Tonic ramping neurons displayed stronger ramping to the 50% reward probability CS that was riskier ($p < 0.01$, time window: 1000ms to 1500ms after fractal onset, sign rank test). (**F**) Reward-timing-uncertain CS responses and post-outcome reward-timing signals. Mean neuronal activity of the tonic ramping neurons (**top**) and phasic bursting neurons (**bottom**). Activity from all trials is shown before 1.5 s, and only from trials in which rewards were not delivered at 1.5 s is shown after 1.5 s. After reward was not delivered at 1.5 s, tonic ramping neurons displayed anticipatory ramping activity for the 4.5 s reward (average ramping across all 3 conditions was related to time: Spearman's rank correlation, $rho=0.32$, $p<0.01$; time window of analyses is indicated by the linear fit over laid on the neuronal activity. Phasic neurons' activity is not significantly related to time (same analyses as above; $p>0.5$). In D-F, neurons with at least 2 trials for each condition (e.g. delivery versus omission) are shown.



Supplemental Figure 2.4. Tonic ramping neurons' activity in the object sequence task. (**A**) Average activity of ramping neurons in the object sequence task. Average activity of tonic ramping neurons ramped to the time of novel fractal presentation and to the time of the reward that followed the sequence (Spearman's rank correlations relating time and activity in the time windows indicated by the linear fits over laid on the neuronal activity, rho=0.29, p<0.01, rho=0.3, p<0.01, respectively). (**A-right**) Area under the ROC curve (AUC) for each ramping neuron that assessed the ability of a neuron to discriminate novel versus familiar objects. Red dots - neurons that significantly discriminate novel versus familiar objects (time window: 200 ms to 400 ms). Overall, the neurons' discrimination was not significantly different from chance (sign rank test; p>0.5). (**B-top**) Histogram of single neurons' response differences for novel fractal in Sequence 2 and Sequence 1, there is no significant difference from 0 (sign rank test; p = 0.79). (**B-bottom**) Same convention as Figure 2.5E. Here the object sequence violation also significantly enhanced neurons' responses (sign rank test; p<0.01). (**C**) Given the results in (A), might ramping occur in anticipation of any salient stimulus or in anticipation of a distracting stimulus? To answer the question, we trained Monkey W to participate in a simple Pavlovian procedure in which ~15% of the inter-trial-intervals contained one of three unexpected 150 ms events (air puff punishments, juice rewards, and synchronous deliveries of white noise bursts with screen flashes). If this first event occurred, a second event (of the same type) always occurred after a fixed time interval. As expected BF ramping neurons responded phasically to the first sound/flash event (black trace). However, after this, they did not significantly ramp to the second sound/flash event. In fact, their activity was clearly reduced by the expectation of the second sound/flash relative to baseline. Responses to other events replicated previous work: during double rewards, the neurons responded to the first unexpected reward with a phasic burst; and the same neurons ramped to the time of punishment-delivery and they subsequently responded with a phasic burst to the delivery of the punishment. To clearly visualize the rapid phasic events, here we used 50 ms Gaussian kernel for to generate the spike density functions. Shaded regions represent SEM.



Supplemental Figure 2.5. Behavioral measures of the motivational effects of object novelty. (**A**) To test if monkeys are behaviorally motivated by novelty, we trained Monkeys R and Z on a saccadic task that measured their eagerness to observe a novel visual object. First, a fixation dot appeared in the center of the screen. 0.5 s after the onset of the fixation dot, a visual object fractal appeared either to the right or the left of the fixation dot (angle: 10 degrees). The monkey was required to continue fixating the dot in the center. After 0.35 s the fixation spot disappeared, and the monkey was free to make saccades. Reward was always delivered 3 seconds after the fractal onset. Therefore, the monkeys' saccadic behavior after the fixation spot disappeared did not affect reward delivery. Monkeys experienced four different trial types. The first two types of trials contained a novel (type 1) or 1 of 2 familiar (type 2) visual fractal objects. Two additional trial types (3-4) tested whether the monkeys were motivated by the possibility of viewing a novel fractal. In trial type 3, 1 of 2 familiar objects appeared. After the fixation spot disappeared, if the monkey fixated the familiar object, it was immediately replaced by a novel object. In trial type 4, 1 of 2 familiar objects appeared. If subsequently the monkey fixated this object, it was replaced by 1 of 2 familiar objects. (**B**) After training (8 days for Monkey Z and 5 days for Monkey R), Monkey R (**top**) and Monkey Z (**bottom**) displayed a decrease in target acquisition reaction time (the time from fixation off to the time the eye fixates the peripheral object) during trials in which the peripheral object was novel versus familiar (first versus second bar; rank sum test p value is indicated above the bars), and when a familiar peripheral object was associated with the presentation of a novel object (third versus fourth bar; rank sum test p value is indicated above the bars). Bars indicate the mean of target acquisition times across all sessions, and the single lines are single sessions' means following training.

Chapter 3: Underpinnings of novelty detection in the primate brain [1]

Primates and other animals must detect novel objects. However, the neuronal mechanisms of novelty detection remain unclear. Prominent theories propose that novelty is either derived from the computation of recency or is a form of sensory surprise. Here, we use high-channel electrophysiology in primates to show that in many prefrontal, temporal, and subcortical brain areas object novelty sensitivity is related to both computations of recency (the sensitivity to how long ago a stimulus was experienced) and sensory surprise (violation of predictions about incoming sensory information). Also, importantly, we studied neuronal novelty-to-familiarity transformations during learning across many days and found a diversity of timescales in neurons' learning rates and between-session forgetting rates within and across brain regions that is well suited to support flexible behavior and learning in response to novelty. Our findings show that novelty sensitivity arises on multiple timescales across single neurons due to diverse related computations of sensory surprise and recency, and shed light on the logic and computational underpinnings of novelty detection in the primate brain.

3.1 Introduction

Humans and other primates learn from the world by exploring objects. Behavioral experiments in primates show that novel visual objects – that is, objects they have never seen before – motivate

[1] This chapter is adapted from an unpublished manuscript (accepted by Current Biology) by Kaining Zhang, Ethan S. Bromberg-Martin, Fatih Sogukpinar, Kim Kocher, and Ilya E. Monosov: "Underpinnings of novelty detection in the primate brain"

behavior, for example, by capturing attention and gaze, and promoting the formation of new memories (Tiitinen et al., 1994; Xiang and Brown, 1998; Bogacz et al., 2001a; Anderson et al., 2008; Joshua et al., 2010; Gottlieb et al., 2013; Ghazizadeh et al., 2016a; Jaegle et al., 2019; Zhang et al., 2019; Tapper and Molas, 2020). And yet, despite the importance of novel objects in our daily life, we currently lack an understanding of how novelty selectivity arises in primate brain circuits and lack an algorithmic understanding of biological novelty detection.

Previous studies reported that neurons in many primate brain areas are novelty responsive – that is, they respond differently to novel versus familiar stimuli (Petrides et al., 2002; Ranganath and Rainer, 2003; Kumaran and Maguire, 2007b; Zhang et al., 2019; Ogasawara et al., 2022). However, novel stimuli differ from familiar stimuli in many respects. For instance, novel stimuli are unexpected or surprising, deviate from recent experiences, and motivate behavior (Berlyne, 1950; Berlyne, 1957; Berlyne, 1960; Berlyne, 1970; Bogacz et al., 2001b; Barto et al., 2013; Ogasawara et al., 2022). Such broad and diverse properties of novelty not only highlight that it is critical to understand the neural mechanisms of novelty detection, but also illustrate why it has been challenging to dissociate representations of novelty from other neural signals, particularly in higher-order brain areas.

There are several formal theories and hypothesized algorithms for processing novelty that each suggests related but dissociable mechanisms for novelty detection. They make distinct predictions about the nature of novelty responsive neurons in the brain (Figure 3.1A). The first one conceptualizes novelty as a form of sensory surprise (Kumaran and Maguire, 2007b; Kumaran and Maguire, 2007a; Egner et al., 2010; Barto et al., 2013; Schwartenbeck et al., 2013; Homann et al., 2017; Reichardt et al., 2020) (Figure 3.1A – Model 1). Sensory surprise is a

violation of predictions about incoming sensory information and could be due to the probability of a specific stimulus or the overall sensory statistics of a given context, such as when expected sequences of objects are violated (Barto et al., 2013; Zhang et al., 2019). In this conception, novelty responsive neurons ought to be sensitive to sensory surprises due to errors in prediction about which sensory events occur. A second class of models conceptualize novelty as a recency and/or repetition effect, which is commonly operationally defined as a neural or behavioral sensitivity to how long ago a stimulus was experienced (Fahy et al., 1993; Li et al., 1993; Xiang and Brown, 1998; Bogacz et al., 2001a; Vogels, 2016) (Figure 3.1A – Model 2). While these two processes could be distinct, these processes could also be interdependent and cooperate (Hart and Jacoby, 1973; Bogacz et al., 2001a), particularly if the brain contains circuits with multiple timescales of object memory. Hence, it is possible that novelty selectivity could arise with both sensory surprise and recency computations (Model 3) or that each contributes to novelty computations preferentially in different brain areas. Finally, novelty responses could arise independently of sensory surprise or recency, for example as a categorical signal for 'complete novelty that purely indicates whether or not a stimulus has ever been seen before (Model 4) (Berlyne, 1960; Miljković, 2010; Barto et al., 2013).

Human studies have examined how novelty, recency, and surprise modulate blood-oxygen level dependent signals (BOLD), as well as other signals that can be acquired non-invasively (Law et al., 2005; Strange et al., 2005; Dudukovic and Wagner, 2007; Wessel et al., 2012; Schomaker and Meeter, 2015; Kafkas and Montaldi, 2018; Utzerath et al., 2018). They studied novelty and recency, or novelty and surprise, but not all three, and, most importantly, could not determine whether these key variables were represented by the same group of neurons or by entirely different neurons within a given voxel or brain area.

We set out to (*i*) test the relationship between novelty, recency, and different forms of sensory surprise and (*ii*) explore the nature and timescales of novelty representations in the activity of single neurons. To do this, we implanted two monkeys with semi-chronic high channel count arrays and recorded neurons across temporal cortex, amygdala, hippocampus, basal ganglia, and the prefrontal cortices while monkeys participated in unsupervised learning object viewing procedures that assessed the relationship of single neurons' object novelty responses with recency and sensory surprise, and dissociated novelty responses from reward value and uncertainty related computations.

Our findings show that novelty sensitivity is heavily intertwined with computations of sensory surprise and recency in single neurons and operates over diverse timescales both within and across brain areas, shedding light on the logic and computational underpinnings of novelty detection in the primate brain. We suggest that novelty selectivity may be constructed by including constituent elements such as sensory surprise and recency.

3.2 Materials and Methods

3.2.1 General procedures

Two adult male rhesus monkeys (S and L; *Macaca mulatta*) were used for the electrophysiology experiments. All procedures conformed to the Guide for the Care and Use of Laboratory Animals and were approved by the Washington University Institutional Animal Care and Use Committee. A plastic head holder and plastic recording chamber were fixed to the skull under general anesthesia and sterile surgical conditions. For each monkey, we implanted semi-chronic high channel count recording drives (LS124; Gray Matter). To aim these micro drives, we first acquired 3T magnetic resonance images of the monkeys' brain. We used these MRIs to aim the

two micro drives towards the regions of interest, including the prefrontal cortex and the temporal cortex. We then attached MRI compatible chambers to the skull using MRI compatible ceramic screws (Thomas). After the animals recovered, we performed MRI with fiducials such that we could estimate and reconstruct the path of each electrode (Daye et al., 2013; Ledbetter et al., 2016a; Dotson et al., 2017; Dotson et al., 2018; White et al., 2019). Next, we implanted both animals with 124-channel micro drives. These are detailed here: https://www.graymatter-research.com/documentation-manuals. Following craniotomy, we sealed the chamber and used a port to assess whether bacterial growth occurred. Following this safety precaution, we implanted the recording drives containing the electrodes and lowered all channels immediately beyond the dura. In this way, we minimized the impact of post-op dura thickening on the electrode impedance and trajectory. Data from electrode-channels were included in the study if (1) post-op CT images showed that the electrodes were in the brain and were following a trajectory that could be reconstructed, (2) if the electrode-channel produced single units during the history of the array neuronal recordings, and (3) if the post-op impedance was >0.2MΩ or single units were observed. This approach produced 108/124 channels in Monkey L and 124/124 channels in Monkey S. A key difference in success was due to the use of glass coated electrodes (Alpha Omega) in Monkey S versus thinner epoxy electrodes in Monkey L (FHC). The semi chronic drive contained electrodes with 1.5mm spacing. Signal acquisition (including amplification and filtering) was performed using Plexon 40kHz recording system. Action potentials were identified using a template matching based algorithm to sort the data, and then we minimized cluster over-splitting, removed artifacts, and selected isolated clusters. All recording and reconstruction procedures are as in (Ogasawara et al., 2022).

3.2.2 Spike sorting

We used the Kilosort algorithm (Pachitariu et al., 2016), specifically, Kilosort2, to sort the action potentials (spikes) from the raw electrophysiological data.

However, the outputs from Kilosort2 tended to be over-split clusters. Thus, we wrote a post-Kilosort algorithm to auto-merge the clusters, remove the artifacts, and label the clusters that are good for subsequent analyses.

The following describes the main part of the post-Kilosort algorithm:

The output of Kilosort2 included the time points, shapes, cluster labels of the spikes, and the templates of the clusters.

The post-Kilosort algorithm at first tried to get rid of artifacts. Because the electrode sites on the array were far separated, if the activity of a cluster appeared on more than one electrode site, it should be an artifact. The algorithm used Kilosort's templates of the clusters and raw spikes to find these artifacts and exclude them in the subsequent analyses. In addition, if a cluster had too low(<0.05Hz) or too high firing rate(>400Hz), the algorithm would label it as artifact as well.

Next, if there were more than one non-artifact cluster in an electrode site, the algorithm would try to merge them. The first three components of the spike shapes which were given by principal component analysis (PCA), and the nonlinear energy of the spike shapes were used to build the clustering space for merging the spikes.

The algorithm picked two clusters each time and tried to merge them, one cluster called "host cluster", and the other called "guest cluster". The guest cluster would be merged into the host cluster if it met the following criteria:

1) The host cluster did not have good isolation in the clustering space, which was measured by silhouette value (silhouette value>0.3, the range is [-1, 1])

2) The distance which was measured by squared Mahalanobis distance from the host to the guest was small. (Squared distance<9).

3) The inter-spike interval (ISI) violation was small after merging. (The violation index<0.5, the range is [0,1])

If ALL criteria were met, the algorithm would merge the guest cluster into the host cluster and delete the guest cluster's label. The algorithm went through all pairs of non-artifact clusters.

After merging, to decide if a cluster was good enough for the subsequent analysis, the algorithm used the following criteria:

1) The cluster was projected onto the first component of PCA, and the distribution was not bimodal, which was tested by Hartigan's dip test (p-value >0.01, the range is [0,1]).

2) The cluster had good isolation from other clusters, and the inter-spike-interval (ISI) violation was low (EITHER the silhouette value > -0.1 and the violation index <0.5 OR the silhouette value > 0.6 and the violation index <0.7. For the electrode site which only had one cluster, the isolation measurement did not work, we required the violation index<0.6).

3) The average spike shape of the cluster was similar to a typical neuron, this criterion included measuring and restricting the variance and the second derivation of the averaged spike shape.

If ALL criteria were met, the cluster was labeled as good and would be included in the subsequent analyses.

3.2.3 Behavioral tasks

Object viewing procedure.

This behavioral procedure was designed to investigate how novelty and novelty-related events are encoded. In each trial of the task, a fixation point appeared at the center of the screen (~0.5 degrees). The shape of the fixation point indicated the trial type (i.e. each of the four trial types had a distinct fixation point shape). The animal was required to fixate on the fixation point for 400ms to initiate the trial. After that, the monkeys were shown a sequence of three fractals at the center of the screen, during which time the fixation point remained at the center of the screen and animals were required to maintain fixation. Each fractal was shown for 250ms and there was a 250ms inter-fractal-interval between fractal presentations. If the animal broke fixation at any time before the third fractal disappeared, or did not start to fixate on the fixation point within 5s after the point appeared, this was counted as an error, the trial stopped immediately, a sound indicating an error was played, and the same trial started again after the inter-trial interval (ITI, ~5s). If the animal successfully fixated till the end of the 3rd fractal, the screen went blank for a randomized time (200ms-1000ms), then a reward dot, visually distinct from the fixation point dot at the start of the trial, appeared in one of four peripheral positions (~10 degrees above/below/left/right of the center of the screen). The monkeys needed to saccade to this dot for reward (Fig. 1B); if the monkey failed to do so within 5s, the reward dot disappeared, a sound indicating an error was played, and the same trial started again after the ITI.

There were four trial types (Type 1 through Type 4), which respectively occurred with 12.5%, 25%, 25%, and 37.5% probability. Each trial type contained a distinct set of fractals. In type 1 trials, the 2nd fractal was always a novel fractal which was generated at the start of the trial,

while the other two fractals were fixed familiar fractals (i.e. the same two familiar fractals were always used for Trial Type 1). In Type 2 trials, all three fractals were familiar fractals presented in a fixed order. There were 2 possible distinct sets of three familiar fractals. On each trial, one of these two sets was randomly picked to be shown, and each set was always shown in the same fixed sequence (i.e. always A→B→C or D→E→F; except for rare 'sequence violation' trials, explained below). We used Type 1 and Type 2 trials to test whether a neuron responds to the predictable onset of a novel stimulus (2nd fractal on Trial Type 1) vs. predictable onset of a familiar stimulus (2nd fractal on Trial Type 2). In trial type 3, just like trial type 2, there were also two distinct sets of three familiar fractals, and one of these two sets was randomly picked to be shown on each trial. However, unlike trial type 2, each presented fractal was drawn randomly with replacement from the picked set. Thus, on each trial, the three presented fractals were drawn from the same set, but they could occur in a randomized order (e.g. A→B→C, B→A→C, etc.) and could include repeats of the same fractal while omitting other fractals (e.g. A→B→A, B→B→B, etc.). This trial type was designed to study surprise and recency, because unlike trial type 2 each individual fractal could not be fully predicted, and there were variable time durations between each time the animal was exposed to a given fractal. In addition, to create sequence violation events, in trial type 2, with 5% probability in each of the 2nd or 3rd positions in the sequence, the familiar fractal that would have been presented on that trial was replaced with the familiar fractal in the corresponding position of the other set (e.g. A->E->C or A->B->F). Thus during a sequence violation, the violating fractal had an unexpected identity, but all fractals still had their normal positions in the sequence and normal proximity to reward.

In trial type 4, there were 3 types of fractals: always novel fractals, repeating novel fractals and familiar fractals. We used this trial type to study neurons' novelty-familiarity transformation. The

always novel fractals were generated before each trial. The familiar fractals were 4 fractals which were always the same and were exposed to the monkey thousands of times. The repeating novel fractals were slightly different for each animal. In Monkey S's version, 4 repeating novel fractals were generated before the task each day. On the next working day, 2 of those 4 fractals were deleted, while the other 2 were saved up to 5 working days and then deleted. Thus, there were always 12 repeating novel fractals in each session: 4 fractals that were on their 1st day of exposure, 2 that were on their 2nd day, 2 that were on their 3rd day, and so on (Figure 3.1B). On each trial, each fractal in each position of the sequence was picked randomly and independently from these types. Thus, the probability of presenting each fractal was: 1/17 for each of the 12 possible repeating novel fractals; 1/17 for each of the 4 possible familiar fractals; and 1/17 to show an always novel fractal. In Monkey L's version, 4 repeating novel fractals were generated before the task each day and were deleted on the next working day, and another 2 repeating novel fractals were generated on the first day of recording and were replaced after 5 days. Thus, there were always 6 repeating novel fractals in each session: 4 fractals that were on their 1st day, and 2 fractals that could be on either their 1st, 2nd, 3rd, 4th, or 5th day. On each trial, each fractal in each position of the sequence was picked randomly and independently from these types. Thus, the probability of presenting each fractal was: 1/11 for each repeating novel fractal, 1/11 for each familiar fractal, and 1/11 to show an always novel fractal.

Usage of fractals as visual stimuli.

All visual fractals were generated using the same previously described algorithm (Miyashita et al., 1991; Yamamoto et al., 2012; Yasuda et al., 2012; Zhang et al., 2019; Ogasawara et al., 2022). In previous work, monkeys strongly and rapidly discriminated novel fractals from the

familiar fractals (Hikosaka et al., 2013; Ghazizadeh et al., 2016a; Ghazizadeh et al., 2020) and learned to distinguish between hundreds of fractals (e.g. associating different individual fractals with reward or no reward) (Yasuda et al., 2012; Hikosaka et al., 2013; Ghazizadeh et al., 2016b). After this training, they still readily detected that a new fractal is novel and not part of a well-learned set (Hikosaka et al., 2013; Ghazizadeh et al., 2016a). Using an algorithmic procedure for generating stimuli has key advantages over the alternative of drawing from a library of objects (e.g. photographs of objects or scenes), including being able to generate a very large number of novel objects on-demand on a trial by trial basis, and ensuring that all stimuli have similar gross visual properties (e.g. size, degree of radial symmetry, etc.) to minimize the possibility that response differences between conditions could be caused by the stimulus sets containing visual features that just happen to vary with the task variables.

Reward information viewing procedure.

This behavioral procedure was a variant of the information viewing task we previously used investigate how reward and information about reward are encoded in the brain (White and Bromberg-Martin et al., 2019, Nature Communications). On each trial of the task, a fixation point appeared at the center of the screen which the monkey was required to fixate for 300ms to initiate the trial. After the trial was successfully initiated, the fixation point disappeared, and a fractal Cue1 appeared on the screen for 1s. This cue indicated the probability of large reward delivery. Then a fractal Cue2 appeared at the center of Cue1. On informative trials, Cue2 provided information about whether a big reward was going to be delivered. Both cues stayed on the screen for another 1s. Monkeys were required to fixate as long as Cue1 or Cue2 were on the screen. If the monkey broke fixation or did not fixate to start the trial 5s after the fixation point

appeared, this was counted as an error, the trial stopped immediately, a sound indicating an error was played, and the same trial started again after the ITI (2s). After the cues disappeared the reward was delivered, which was always either a large or small amount of juice. Two blocks of trials alternated: In the informative block, Cue2 informed the monkey whether the big reward was going to be delivered, while in the non-informative block Cue2 was randomized and hence provided no new information about the outcome. In both blocks, Cue1 indicated the probability of big juice delivery (0%, 50%, or 100%). In each block, there were two possible Cue1 stimuli for each probability (one of which was randomly chosen to present on each trial), thus there were a total of 12 unique Cue1 fractals). In the informative block, there were 4 possible Cue2 stimuli, 2 indicating big reward and 2 indicating small reward. On each trial, one of the 2 stimuli corresponding to the trial's upcoming reward outcome was randomly chosen to be presented. In the non-informative block, there were 4 distinct possible Cue2 stimuli. On each trial, one of these 4 stimuli was randomly chosen to be presented (and hence conveyed no information about the reward outcome).

3.2.4 Data analyses

For all analyses in the object viewing procedure, unless otherwise stated, each neuron's responses to visual fractal objects were measured as the mean firing rate in the 500ms time window starting from fractal onset. When permutations were used to assess significance or obtain confidence intervals, we permuted 10000 times.

Novelty index was quantified by AUC of ROC (area under the curve of the receiver operating characteristic curve) comparing the neural responses to the novel fractals in the second position of Type 1 trials versus the familiar fractals in the second position in Type 2 trials. We then

subtracted 0.5 from the AUC such that numbers higher than 0 indicated that the neuron had a higher firing rate to the novel fractals than familiar fractals, and multiplied the result by 2 such that the range of the index was [-1, 1]. The significance of this index was tested by a rank sum test (threshold: $p<0.01$).

A novelty responsive neuron was defined as a neuron with a significant novelty index. A novelty-excited neuron was a novelty responsive neuron with its novelty index larger than 0, while a novelty-inhibited neuron was a novelty responsive neuron with its novelty index less than 0. We used analogous definitions for recency responsive neurons, sensory surprise responsive neurons, recency-excited neurons, etc.

Sensory surprise index was quantified by AUC of ROC of each neuron's responses to the familiar fractals in the third place in Type 3 trials versus the familiar fractals in the third place in Type 2 trials (excluding the ~10% fractals in Type 2 trials that were sequence violations, and the non-sequence-violating fractals that were used to calculate the violation index). This compared neural responses for predictable versus unpredicted familiar objects. We subtracted 0.5 from the AUC, so that values higher than 0 indicated that the responses were higher for unexpected fractals versus expected fractals, and multiplied by 2 so that the range of the index was [-1, 1]. The significance of this index was tested by a rank sum test (threshold: $p<0.01$). To further eliminate the effect of recency, we performed a 1-way ANOVA analysis (MATLAB) to measure the effect on each neuron's firing rate of whether the fractals occurred within the same trial (recent) or a previous trial (nonrecent), then subtracted the effect of recency from each neuron's responses before repeating the above ROC analysis.

Object recency index was quantified using Type 3 trials. We categorized each fractal's presentation based on whether the most recent presentation of the same fractal had occurred within the same trial (recent) or a previous trial (nonrecent). The recency sensitivity index of each neuron was quantified by AUC of ROC comparing responses for nonrecent versus recent objects. We subtracted 0.5 from the results, so that values higher than 0 indicated that the neuron had higher firing rate to nonrecent fractals than recent fractals, and multiplied by 2 so that the range of the index was [-1, 1]. For this analysis we only used object responses during the 2nd and 3rd position in the Type 3 trial sequence (so that it was possible for the object to be either recent or nonrecent). To remove sequence position effects, we subsampled the data before performing ROC analysis so that there were an equal number of recent and not recent objects at each sequence position (maximizing the contrast between conditions whenever possible by choosing the subset of nonrecent objects that were 'least recent', i.e. which had the longest time duration since their last presentation). To further eliminate the effects of sequence position and object selectivity, above and beyond the position matching procedure described above, we performed a 2-way ANOVA analysis (MATLAB) on position and object identity, then subtracted the effect of position and object selectivity from each neuron's responses before performing the above ROC analysis. The significance of this index was tested by a rank sum test (threshold: $p<0.01$).

Sequence violation index was measured using the sequence violation trials described above. We calculated the AUCs of ROC of the neuron's firing rate to the familiar sequence-violating fractal versus the familiar non-sequence-violating fractal in the 2nd place and 3rd place respectively. We then averaged these two AUCs, and subtracted 0.5 such that numbers above 0 indicated that the neuron had higher firing rate to the sequence violated fractals than the normal fractals, and

multiplied the result by 2 so that the range of the index was [-1, 1]. The significance of this index was tested by a permutation test (threshold: $p<0.01$).

Reward value index. In the reward information viewing procedure, we quantified sensitivity to changes in reward value signaled by visual objects by an AUC of ROC that compared neuronal responses to the 100% reward versus 0% reward trials (pooling informative and non-informative trials) in the last 0.5s epoch before the reward was delivered. We then subtracted 0.5 from the result of the ROC analysis such that values higher than 0 indicated that the neuron had a higher firing rate to higher reward cue than lower reward cue, and multiplied the result by 2 so that the range of the index was [-1, 1]. The significance of this index was tested by a rank sum test (threshold: $p<0.01$).

Information anticipation index. This index was adapted from the informative cue anticipation index used previously to measure how strongly a neuron anticipated the receipt of informative visual cues to resolve uncertainty about upcoming rewards (White et al., 2019; Jezzini et al., 2021). It was defined as the difference between the magnitudes of neuronal uncertainty signals during informative versus non-informative trials, where uncertainty signal was defined as the AUC of ROC comparing neural activity on trials where Cue1 indicated an uncertain reward outcome (50% big) vs. a certain reward outcome (either 100% big or 0% big). In essence, this index measured how strongly a neuron anticipated information to resolve uncertainty (White et al., 2019; Jezzini et al., 2021). The range of the index is [-1, 1]. The significance of this index was tested by a permutation test (threshold: $p<0.01$).

Normalization of firing rate in the learning analysis. We z-scored each neuron's firing rates by the mean and standard derivation of the firing rates from all types of trials. and then averaged

the z-scored firing rates to always novel fractals, familiar fractals, repeated novel fractals on day 1, and repeated novel fractals on days 2+, separately for each presentation within the current day. Then, for each separate presentation number in the session, we rescaled the firing rates so that the averaged normalized firing rate was 0 for always familiar fractals and 1 for novel fractals. The error bars of the normalized firing rate for repeated novel fractals were the standard errors of the mean (computed using bootstrapping, n=10000 bootstraps). This analysis controls for repetition suppression because activity for repeating novel fractals is normalized to be relative to activity for always familiar fractals, and both of these sets of fractals were repeated over the course of the session in exactly analogous manners.

Learning rate analysis. The learning rate at the nth presentation in Figure 3.5B lower panel was calculated as follows:

$$\alpha(\mathrm{n}) = \frac{\mathrm{R(n)} - \mathrm{R(n+1)}}{\mathrm{R(n)}} \tag{3.1}$$

Where R(*k*) in the equation is the population average normalized firing rate in response to the *k*-th presentation of the fractal during a session. We calculated the learning rate separately for 1st day fractals and 2nd+ day fractals. We then smoothed the learning rate using a 3-apperance bin. The error bars in Figure 3.6B were the bootstrap standard error of the mean (n=10000 bootstraps).

Within day learning index. For each novelty-excited neuron, we compared its object responses during the first 5 and the last 5 presentations of the repeated novel fractals (Figure 3.1). To do this, we needed to quantify the strength of responses to the repeated novel fractals relative to responses to the always novel fractals and the familiar fractals. To accomplish this, we used a

classifier to estimate the approximate posterior probability that each of these responses was evoked by an always novel fractal, given an equal prior probability of it being evoked by either an always novel fractal or a familiar fractal. To classify the firing rates to repeated novel fractals, the classifier used the equation:

$$classifier(R) = \frac{1}{(1 + \exp\left(\frac{a * (R - b)}{\sigma^2}\right))} \tag{3.2}$$

$$a = N - F$$

$$b = \frac{N + F}{2}$$

Where R is the firing rate to the repeated novel fractal, N is the mean firing rate to always novel fractals, F is the mean firing rate to familiar fractals, and σ^2 is the residual variance of the firing rate to novel and familiar fractals. All of these firing rates are the non-normalized firing rates from the individual neuron (in spikes per second) with the effect of sequence position being subtracted.

This classifier gives a result in [0,1]. This can be interpreted as the posterior probability of the response being generated from a novel fractal, if a neuron's firing rates to novel and familiar fractals are Gaussian distributions with different means but the same variance, and both types of fractals are equally likely to have been presented. The within day learning index was defined as the difference between the mean of classification results of the repeated novel fractals at the start of day 1 vs. the end of day 1. The range of this index is [-1, 1]. Note that this analysis controls for repetition suppression because activity for repeating novel fractals is classified relative to

activity for always familiar fractals, and both of these sets of fractals were repeated over the course of the session in exactly analogous manners.

Across day forgetting index. We used the same classifier as the within day learning index. The across day forgetting index was defined as the difference between the mean of classification results of repeating novel fractals at the beginning of the second and subsequent days vs. at the end of the first day. In the calculation, non-overlapping sets of fractals at the end of the first day were used in the calculation of within day learning index and across day forgetting index, such that the two measurements were independent. The range of this index is [-1, 1]. For comparison of brain areas' learning and forgetting indices (Figure 3.7C) we included brain areas with at least n=20 novelty excited neurons.

Control for Figure 3.7B. We controlled for session-to-session variability in learning, which might contribute to the correlation of within day learning index and across day forgetting index. In other words, could it be that within every single session all neurons learn and forget at the same rate as each other, but some sessions are 'fast' and other sessions are 'slow'. If so, the apparent differences in learning speeds across neurons could arise from differences in learning across sessions, not neurons *per se*. To control for this possibility, we used the following analysis. For each day, if the recording session had >= 5 novelty-excited neurons, the session was included. For each session we calculated the averaged within day learning index over all of these neurons and then subtracted it from each neuron's individual within day learning index. We did the analogous procedure for the across day forgetting index. Lastly, we calculated the correlation of the subtracted indices.

Pupil diameter was obtained with an infrared video camera (Eyelink, SR Research). To quantify pupil's response to novel, surprising, and nonrecent fractals (Figure 3.2A), we used the following procedure. First, we z-scored the pupil diameter on a trial-by-trial basis (time window from the start of the first fractal until 250ms after the third fractal disappeared). Next, for each fractal within the trial, we subtracted the baseline z-scored pupil response (response in the time window [-80, 20] relative to fractal onset). We then averaged the pupil response in the time window [0ms 500ms] relative to fractal onset. Lastly, we calculated the same novelty, sensory surprise, and recency indices as we used for the neurons and then multiplied by -1, such that the index was positive if the pupil contracted more to novel/surprising/nonrecent objects.

Correlation analyses. In Figure 3.3A and Supplemental Figures, the novelty responsive neuron group contained all novelty responsive neurons. In order to combine the results of novelty excited neurons and novelty inhibited neurons together, we flipped the signs of the all the indices of neurons whose novelty index is negative, while the indices of neurons whose novelty index is positive remained unchanged. In supplemental figures, all indices of neurons were in their original signs. In Figure 3.3C and supplemental figures, the bar plots were generated by binning the novelty index (x axis) into three even parts, [0,1/6], [1/6, 1/3], [1/3, 1/2], and we calculated the mean and the standard error of the mean of the indices in y axis of the neurons in each bin. For the linear fitting we used least squares regression. For all correlation analysis, we used Spearman's rank correlation unless stating specifically.

Classifier analyses. In Figure 3.3B and Supplemental Figures we used the activity of all neurons to train and test two support vector machine (SVM) classifiers. One classifier was trained to decode sensory surprise and then was tested to determine whether it could be used to decode

novelty. The other classifier was trained to decode recency and then tested to decode novelty. The training set for sensory surprise was the combination of sensory-surprising fractals (label as 1, they were the familiar fractals in the third position in Type 3 trials) and non-sensory-surprising fractals (label as 0, they were the familiar fractals in the third position in Type 2 trials). The training set for recency was the combination of nonrecent fractals (label as 1, the fractals that had occurred in the previous trial in Type 3 trials) and recent fractals (label as 0, the fractals that had occurred within the same trial in Type 3 trials). The test set for the two classifiers was the same. They included novel fractals (label as 1, novel fractals in Type 1 trials) vs. familiar fractals (label as 0, familiar fractals in the second position in Type 2 trials). To avoid introducing base rate biases into the classifier, in both training and testing set, the numbers of fractals with labels 1 and 0 were balanced by subsampling. We applied this classifier analysis separately for each session, computed the percentage of objects that were classified correctly, and then averaged this percentage across sessions.

Cross-validation. For activity plots in supplemental figures, we first separated the sessions into even trials and odd trials. Then we calculated the relevant indices using odd trials and used this to select neural activity from the even trials for those neurons whose indices in the odd trials were significant ($p<0.01$) and excited (index>0) or inhibited (index<0). Also, we calculated the indices using even trials and used this to select neural activity from the odd trials of the neurons whose indices in the even trials were significant ($p<0.01$) and excited (index>0) or inhibited (index<0). This ensured that the analysis was cross-validated, i.e. each piece of data from a neuron was selected to be used for this analysis on the basis of a separate, independent set of data from the same neuron. Finally, we plotted the average z-scored activity from the selected odd and/or even trials of the neurons. (If an individual neuron's odd and even trials were both

selected, then both were contributed to this analysis). All PSTHs were smoothed by a Gaussian kernel (SD = 50ms). The p value in the PSTHs plots were rank-sum tests of the average of the two PSTHs in the target window ([0ms, 500ms] relative to the onset of fractals in object viewing procedure.)

Noise correlation analysis. In each session, we calculated the noise correlation (Pearson's correlation) for each pair of novelty responsive neurons responding to the novel fractals in Type 1 trials. We averaged the correlations across all pairs within the session, and then averaged across sessions. We did the same process on familiar fractals in Type 2 trials in the second position in the object sequence. We also did the same calculations using non-novelty responsive neurons.

Noise variance analysis. This analysis only included sessions with at least 5 novelty responsive neurons. In each session we defined an *n*-dimensional space where each dimension was the firing rate of one of the *n* novelty responsive neurons, and hence the response to each individual fractal presentation could be represented as a point in that space. We defined the novelty axis as a unit vector pointing from the mean of the points representing familiar fractal presentations (fractal presentations in the second position in the sequence in Type 2 trials) to the mean of the points representing novel fractal presentations (fractal presentations in the second position in the sequence in Type 1 trials) (Supplemental Figure 3.6). We defined random axes as unit vectors drawn randomly from a uniform distribution on the unit sphere. We then computed the ratio of the mean neural response variance projected onto the novelty axis vs. random axes, as follows. We randomly chose 5 individual fractal presentations to represent each of the following three conditions. Condition 1: 5 different novel fractals (from the second position in Type 1 trials).

Condition 2: 5 individual presentations of different familiar fractals (from the second position in Type 3 trials). Condition 3: 5 repeated presentations of the same familiar fractal (from the second position in Type 3 trials, using only the remaining familiar fractal from Type 3 trials that was not included in Condition 2). We also randomly chose one random axis. For each condition, we then computed the variance of its 5 individual neural responses when they were projected onto the novelty axis, and when projected onto the random axis. We repeated this process 10,000 times, using different random selections of individual fractal presentations for each condition and a different random axis. We then computed the ratio of response variances for each condition as the mean of the 10,000 variances along the novelty axis divided by the mean of the 10,000 variances along the random axes. This produced one response variance ratio for each of the three conditions in each session. We then averaged these ratios within each condition over sessions, and tested the difference between the conditions using signed-rank tests.

Stability of object selectivity across sessions. We tested whether object selectivity of single neurons changed during novelty-familiarity transformations. In each session, we divided the learning fractals into three groups chronologically, and used the early group (start of the session) and the last group (end of the session) separately to measure object selectivity for each neuron. To obtain an object selectivity index we performed a 2-way ANOVA analysis on sequence position and object identity, and obtained a measure of variance explained by object identity(Var_{id}), variance explained by sequence position ($Var_{position}$) and residual variance (Var_{res}).

$$object\ selectivity\ index\ = \frac{Var_{id}}{Var_{id} + Var_{res}} \tag{3.3}$$

In novelty-responsive neurons, we found that there was no significant difference in their object selectivity between the start (early group) and the end of the sessions (last group) ($p = 0.24$, signed-rank test).

Hierarchical clustering of brain areas. We performed hierarchical clustering of brain areas based on the strength of their sensory surprise and recency effects relative to their novelty effects, using the following procedure. First, for each neuron we converted its novelty, sensory surprise, and recency indexes into unsigned "absolute" indexes to represent the overall strength of its coding, by multiplying each index by -1 if it had a negative sign while leaving it unchanged if it had a positive sign. Then, for each brain area, we computed the mean of each these three absolute indexes across its neurons. Finally, in order to measure sensory surprise and recency relative to novelty, we normalized the mean absolute sensory surprise and recency indexes for each area by dividing them by that area's mean absolute novelty index. Thus each area was represented as a point in a two dimensional space, defined by its normalized mean absolute indexes for coding of sensory surprise and recency. We then used hierarchical clustering to cluster the areas based on their Euclidean distance in that space and using the unweighted pair group method with arithmetic mean (UPGMA)

3.3 Results

3.3.1 A passive object-viewing behavior procedure is used to detect the novelty, sensory surprise, and recency responses in the macaques' brain.

On each trial, the monkey was shown a sequence of three fractal visual objects. The objects in these sequences did not possess instrumental value and did not affect reward rate or magnitude (Materials and Methods). Monkeys obtained reward after successfully observing each object

sequence and then participating in a distinct instrumental behavior that consisted of making an eye movement to a dot that appeared at one of four possible locations on the screen (Figure 3.1B, top right).

To observe the relationship of novelty with sensory surprise and recency, this procedure contained several trial types which included distinct object sequences designed to dissociate these factors (Materials and Methods). The trial type was cued to the animal at the start of the trial by the shape of the fixation point (Figure 3.1). In Type 1 trials, the monkeys experienced a sequential presentation of three objects, in which the second object was always novel and the other objects were familiar and fully predictable (Figure 3.1B). The novel objects were never before seen because they were generated on a trial-by-trial basis using a new random seed on each trial (using a previously established algorithm (Miyashita et al., 1991; Yamamoto et al., 2012; Yasuda et al., 2012; Zhang et al., 2019; Ogasawara et al., 2022). In Type 2 trials, monkeys experienced other distinct fractal visual objects that were all highly familiar. For these trials, we used two sets of objects to control for single neurons' object sensitivities (Figure 3.1B, Materials and Methods), and following the presentation of the first fractal, the remaining objects in the sequence were predictable (Figure 3.1B). Hence, the variability or entropy of which object would be presented was relatively low. We defined novelty responsive neurons as those that responded differentially to the second objects in Type 1 versus Type 2 trials (Zhang et al., 2019). Importantly, this design ensured that it was highly predictable whether the second object in these sequences would be novel or familiar, so that neural novelty responses could not be attributed to the novelty simply being more unpredictable or surprising than familiarity.

Importantly, during the same recording session, we also measured neuronal sensitivities to object recency and sensory surprise (Materials and Methods). This was accomplished with Type 3 trials that contained three objects that were each drawn from a familiar set of three fractals, but were drawn in a random sequence with replacement (Figure 3.1B). Thus, following the presentation of the first fractal in Type 3 trials, the monkey could predict which set of fractals the remaining two objects would be drawn from, but could not fully predict their specific object identities. Hence, variability or entropy of which object would be presented was relatively high. By comparing Type 2 and Type 3 trials, we measured neural sensitivity to sensory surprise - that is, responses that were attributable to the presentation of an object whose identity and sensory features were predictable vs. unpredictable. Furthermore, neural sensitivity to recency was assessed by comparing responses to familiar objects during Type 3 trials that were more or less recently seen (Figure 3.1B, Materials and Methods; (Xiang and Brown, 1998)). Hence, in Type 1-3 trials, for each neuron, we obtained independent measures of sensitivity for novelty, surprise, and recency. Importantly, these measurements were independent of each other (Materials and Methods). Thus, any relationship between these measures reflects a relationship between how neurons generate their responses to these variables (such as is hypothesized by the models in Figure 3.1A). We also used a distinct set of trials to study learning and the timescales of novelty-to-familiarity transformations (Type 4; we return to them later).

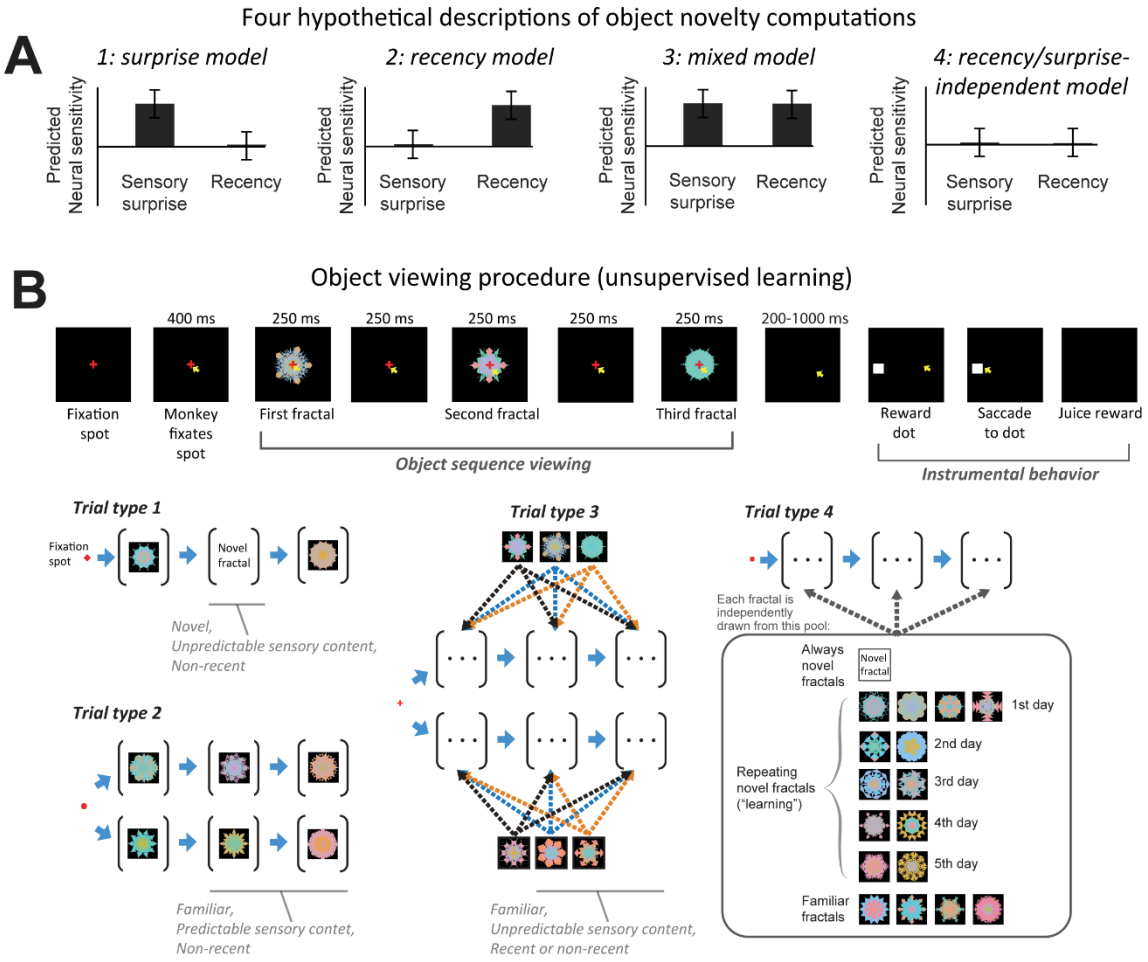


Figure 3.1. Models of object novelty computations and object viewing procedure. (**A**) Four theoretical descriptions of the mechanisms of object novelty detection predict different patterns of neural responses for object recency and sensory surprise. (**B**) Object viewing procedure (**top**) and four trial types (**bottom**).

Monkeys showed behavioral evidence that they understood the procedure. Because we used an unsupervised object viewing procedure that required the monkeys to fixate during object sequences, we were also able to measure pupillary responses to novelty, recency, and sensory surprise. We found that both animals were sensitive to the task because their pupillary constrictions reliably changed as a function of these variables (Figure 3.2A. Monkey L: novelty index, $p<0.0001$; sensory surprise index, $p<0.0001$; recency index, $p<0.0001$. Monkey S: novelty index, $p<0.0001$; sensory surprise index, $p=0.57$; recency index, $p<0.0001$). The constriction of pupils to expected novel stimuli in both monkeys replicates previous data in humans (Võ et al., 2008; Kafkas and Montaldi, 2015; Kafkas and Montaldi, 2018). In addition, Monkeys had faster reaction times to initiate Type 1 trials, which contained a novel fractal, than Type 2 trials (Supplemental Figure 3.1A, Type 1 vs. Type 2, $p = 0.00014$). This novelty seeking behavior is consistent with previous findings (Ogasawara et al., 2022). Monkeys also had faster reaction times to initiate Type 3 trials which contained objects that were less predictable than Type 2 trials (Type 2 vs. Type 3, $p<0.0001$).

In our neuronal data, we found a strong relationship between the encoding of novelty and both surprise and recency, consistent with model 3 (Figure 3.2B-C). An example neuron recorded in the amygdala is shown in Figure 3.2B. This neuron robustly discriminated among novel and familiar objects by displaying greater excitation to predicted novel objects than to predicted familiar objects (Figure 3.2B – left). The neuron was also responsive to surprise: it was relatively more excited by unpredicted versus predicted familiar stimuli (Figure 3.2B – middle). The neuron was also responsive to recency: it was more excited by objects that were presented relatively less recently (Figure 3.2B – right). Thus, this neuron was excited by all three types of objects – novel objects, surprising objects, and less-recent objects. We quantified this result by

computing an index of each neuron's sensitivity for each of these types of objects (Figure 3.2C; Materials and Methods). Indices >0 indicate a preference for novelty, sensory surprise, or less-recent objects respectively. The results of the example neuron (Figure 3.2C) strongly resembled model 3 (Figure 3.1A).

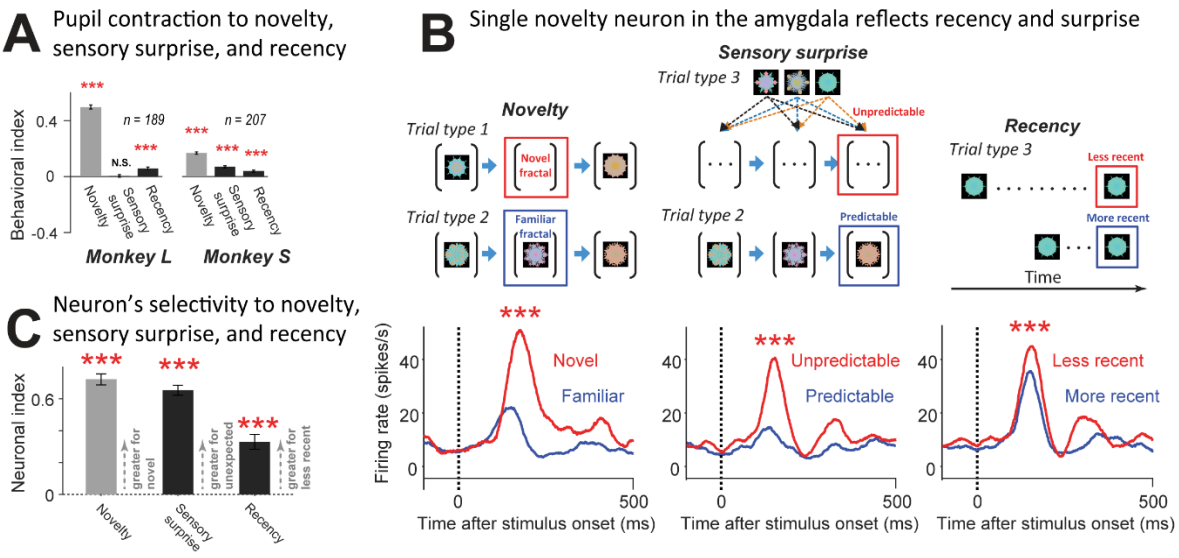


Figure 3.2. Pupillary and neural signatures of novelty coding. (**A**) Pupillary responses of monkeys L and S. Error bars are SE. (**B**) A single neuron in the amygdala is sensitive to novelty, sensory surprise, and recency. PSTHs were smoothed using a gaussian kernel (SD = 50ms). (**C**) Sensitivity indices based on the activity in (**B**). *** denotes p<0.001, error bar indicates bootstrap SE. The derivation of the indices is detailed in Materials and Methods and conceptually shown by the cartoon in (**B**). Blue and red boxes in the cartoon mark the conditions corresponding to the spike density functions in **B** that were used to derive the sensitivity indices.

3.3.2 Novelty sensitivity correlates with both the sensitivities to sensory surprise and recency across neurons and across brain areas

We observed a similar pattern of results across the population of novelty responsive neurons across many brain areas (n=2234). Novelty responsive neurons displayed significant sensitivity to sensory surprise and recency (mean of sensory surprise index = 0.0468, mean of recency index = 0.0263, both greater than 0, $p<0.0001$, signed-rank tests; Figure 3.3A). Crucially, those effects were much greater in novelty neurons versus all other recorded neurons (Figure 3.3A; $p<0.0001$, rank-sum test; Supplemental Figure 3.1 for each animal). In fact, as further confirmation of this result, we found that object novelty could be decoded purely based on neural tuning to surprise and recency. That is, we trained a classifier on all neurons in each session to decode whether the fractals were sensory surprising and another classifier to decode whether the fractals were nonrecent. We found that the two classifiers were both able to decode object novelty significantly above chance (Figure 3.3B, Supplemental Figure 3.1D, E). Sensory surprise classifiers, $p<0.0001$; recency classifiers, $p<0.0001$). Furthermore, on a neuron-by-neuron level, for novelty excited neurons, the magnitude of sensitivity to novelty was correlated with the magnitude of sensitivity to sensory surprise (Figure 3.3C-left) and recency (Figure 3.3C-right; Supplemental Figure 3.2). This data indicates that novelty detection is strongly associated with encoding of sensory surprise and recency. We also verified that these effects were not produced by other factors like object selectivity and sequence position effects or by any potential statistical dependencies between the indexes (Supplemental Figure 3.2).

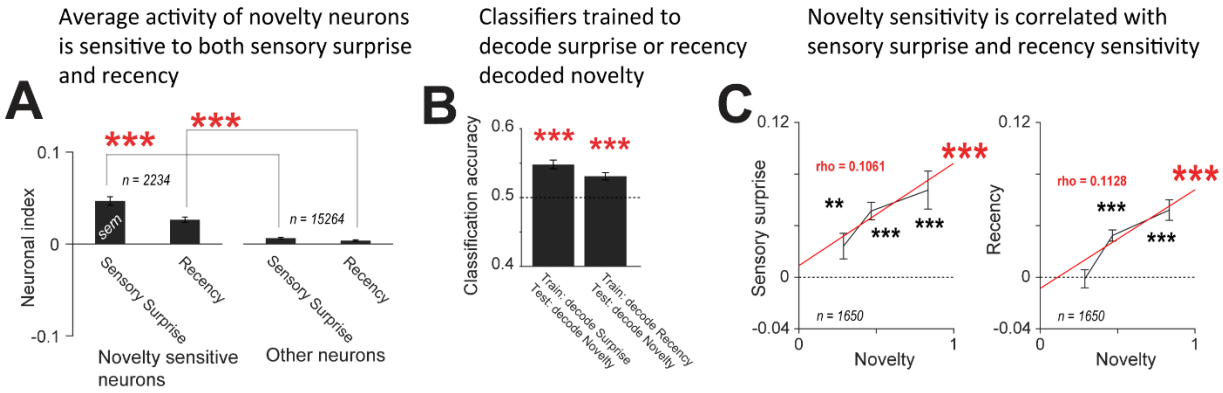


Figure 3.3. Novelty neurons are sensitive to sensory surprise and recency. (**A**) Novelty responsive neurons displayed particularly strong sensitivity to sensory surprise and recency. Sensory surprise and recency population indices are shown for novelty responsive neurons (**left**) and all other neurons (**right**). (**B**) Two classifiers' performance (left and right) at classifying objects as novel versus familiar. These classifiers were trained on responses to familiar fractals to discriminate sensory surprising fractals (left), and recent vs. nonrecent fractals (right). (**Materials and Methods**) (**C**) The magnitude of novelty-excited neurons' novelty sensitivity was correlated with the magnitude of their sensitivity to sensory surprise (**left**) and recency (**right).**

We found that novelty sensitivity was not uniformly distributed throughout single neurons in the primate brain. Some brain areas, in particular the anterior ventral medial temporal cortex (AVMTC), and areas that it is interconnected with, such as area 46v, basal forebrain, and the amygdala, were preferentially enriched in novelty responsive neurons (Figure 3.4A). Consistent with Figure 3.3A-C, we found that there was an across-region relationship between novelty, recency, and sensory surprise – that is, on average, regions that were preferentially enriched with novelty responsive neurons were also enriched with neurons that were responsive to recency and sensory surprise (significantly higher percentage of sensory surprise responsive neurons ($p < 0.0001$, permutation test) and recency responsive neurons ($p < 0.0001$, permutation test) when comparing the $1/5^{th}$ of areas with the highest percentage of novelty responsive neurons to the $1/5^{th}$ of areas with the lowest percentage of novelty responsive neurons; Figure 3.4A, left).

This finding raised the question of whether the relationship between novelty and sensory surprise and recency is only relevant to a small number of brain areas that are most enriched with novelty responsive neurons, or whether this relationship is a general feature common to brain areas involved in novelty processing. Our data indicates that the latter is the case: there were remarkably consistent positive correlations between novelty and both sensory surprise and recency across almost all of the recorded brain areas (Figure 3.4A-right). This suggests that the results in Figure 3.3C generalize across most brain areas that we targeted in this study. In sum, these data show that that novelty is linked measurements of sensory surprise and recency in many brain areas.

This finding also raised the question of whether novelty processing is linked to other types of sensory surprises as well. For example, we recently showed that novelty responsive neurons in

the basal forebrain are sensitive to violations in object sequences, responding to when a familiar object from one sequence unexpectedly appears in another (Zhang et al., 2019). To replicate this and test whether it holds across the many brain areas recorded here, we introduced object sequence violations on a small fraction of Type 2 trials (~10%, Materials and Methods). This produced a remarkably similar pattern of results: novelty responsive neurons were significantly sensitive to sequence violations, and their sensitivity to novelty and sequence violations were positively correlated, particularly in the basal forebrain (Supplemental Figure 3.3). This indicates that the results in Figure 3.3A, B can generalize to other types of sensory surprises, and may highlight both specificity and importance of the basal forebrain in novelty and surprise computations across its cortical and subcortical projection targets (Turchi et al., 2018; Monosov, 2020).

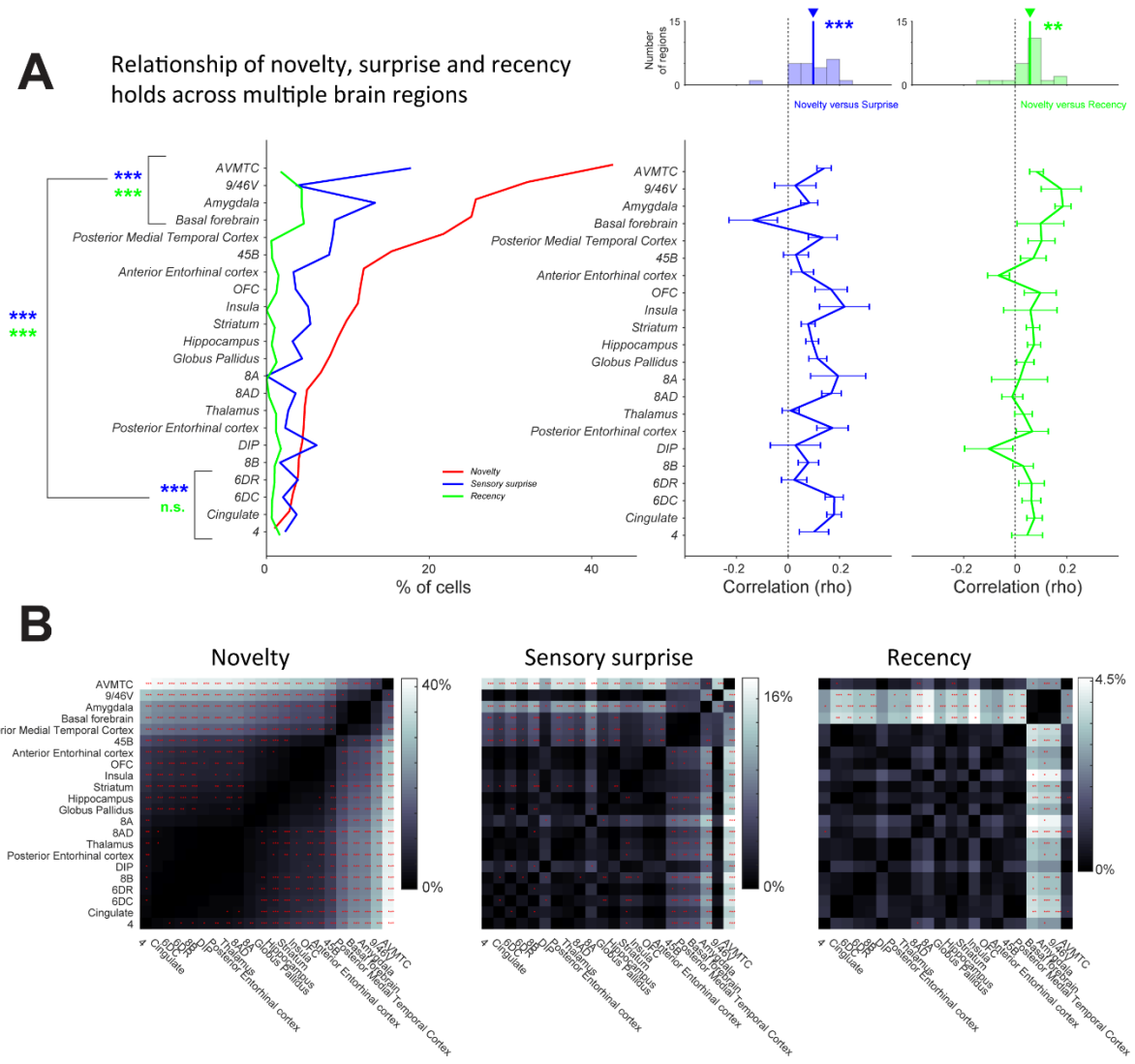


Figure 3.4. Novelty neurons sensitivities to sensory surprise and recency in different brain regions. (**A**) Novelty, sensory surprise, and recency computations often co-occur across the brain. % of neurons significantly responsive to novelty (red), sensory surprise (blue), and recency (green) shown across brain areas that were rank ordered by the percentage of neurons responsive to novelty (**left**; x-axis). The top four brain areas had more neurons than would be expected by chance that were responsive to recency and surprise (colored asterisks), but the bottom four did not (binomial test). The differences in ratios among them was significant ($p<0.0001$, permutation test). Within the brain areas, novelty vs. sensory surprise (**middle**) and novelty vs. recency (**right**) are positively correlated in most brain areas (**top histograms,** signed-rank test relative to 0, *, **, ***, indicate $p < 0.05, 0.01, 0.001$). Error bars indicate SE obtained through a bootstrapping procedure. (**B**) Pairwise comparison between brain areas of their percentage of novelty responsive neurons (left panel), sensory surprise responsive neurons (middle panel), and recency responsive neurons (right panel). The colors in the matrix represent the absolute difference in percentages of responsive neurons, and the red asterisks indicate

whether the differences are significant (Fisher's exact test, *, **, ***, indicate $p < 0.05$, 0.01, 0.001)

Importantly, however, when we interrogated the fine structure of this large-scale brain network we found evidence that novelty, sensory surprise, and recency are not simply treated identically to each other in the brain. Rather, certain areas may have special roles in each of these computations. To test this, we computed a distance matrix between brain areas based on their differences in the percentage of cells with significant novelty, sensory surprise, and recency coding (Figure 3.4B). These three distance matrices were highly correlated with each other, with the same cluster of 4-5 areas standing out from the rest (Novelty vs. Recency, $\rho = 0.502$, $p<0.0001$, Novelty vs. Sensory surprise, $\rho = 0.585$, $p<0.0001$, Sensory surprise vs. Recency, $\rho = 0.327$, $p<0.0001$, Spearman's rank correlation) However, within this cluster there were key differences, suggestive of roles in different stages of novelty-related computations. AVMTC was highly enriched in novelty and sensory surprise coding but less so in recency; 9/46V was enriched in novelty and recency but less so in sensory surprise; while amygdala and basal forebrain integrated all three. In addition, recency and surprise sensitivity were significantly positively correlated across all neurons (rho=0.07, $p<0.001$; animal S, rho=0.1154, $p<0.001$; animal L, rho=0.025, $p=0.0255$) but not always across novelty-excited neurons (rho=0.04, $p=0.0896$; animal S, rho=0.25, $p<0.001$; animal L, rho=-0.125, $p<0.001$). Thus, our data suggest that specific subpopulations of novelty responsive neurons are especially sensitive to either surprise, recency, or both.

To further explore this notion, we performed hierarchical clustering of the brain areas based on their discrimination of novelty, sensory surprise, and recency (Supplemental Figure 3.3). The hierarchical clustering visualized clusters of brain areas in a rank that roughly matches what is

known about their anatomical connections, such as the major connections between basal forebrain, amygdala, and the temporal cortex (Mesulam et al., 1983; Russchen et al., 1985; Stefanacci et al., 1996; Suzuki, 1996; Cheng et al., 1997; Saunders et al., 2005; Monosov et al., 2015). Thus, in addition to the general phenomenon of correlated codes, the fine structure of these correlations across large-scale brain networks indicates that specific areas are best suited for specific forms or stages of novelty-related computations.

3.3.3 Neurons' novelty sensitivities have weaker or no correlation with the sensitivities to reward-related task variables compared with recency and sensory surprise.

Another important question was whether these correlated codes could have been induced not by common computations but by common changes in arousal in response to novelty, surprise, and recency. In our data, this was unlikely because while neural populations recorded in each monkey displayed similar patterns as Figure 3.3 (Supplemental Figure 3.1), their pupillary responses to sensory surprise differed (Figure 3.2A) suggesting that the main population-level results could not be simply a reflection of pupil-indexed arousal. As a further test of this issue, we recorded the activity of many of the same neurons while manipulating arousal in a reward information viewing procedure that measured how novelty sensitive neurons anticipated and responded to information about future rewards, Particularly, in this procedure, animals observed visual objects that indicated changes in reward value and uncertainty (Figure 3.5A), which are known to strongly activate neural populations that regulate motivated behavior (White et al., 2019; Jezzini et al., 2021).

We found that our sampled neural populations were significantly enriched in neurons that were responsive to two important reward statistics known to drive arousal: (i) many neurons signaled

reward values by discriminating among high value and low value objects and (ii) many neurons anticipated the receipt of information to resolve reward uncertainty, by activating when reward was uncertain, and increasing this activity before an informative cue would appear to indicate the reward outcome (Supplemental Figure 3.4A,B). Neuronal sensitivity to these two variables was not correlated with neuronal sensitivity to novelty, among novelty excited neurons (Figure 3.5B). This is consistent with a previous study that found a similar result using functional magnetic resonance imaging in many brain regions ((Ghazizadeh et al., 2020), but note that in brain regions involved in controlling the deployment of spatial attention and gaze control, they found that there was indeed a correlation between novelty and reward). The main point of these analyses was that neuronal sensitivity to novelty was generally more strongly correlated with recency and with sensory surprise, than it was with reward value and with information anticipation (novelty and recency vs. novelty and reward value, $p=0.0008$; novelty and recency vs. novelty and reward information $p=0.052$; novelty and sensory surprise vs. novelty and reward value, $p=0.0114$; novelty and sensory surprise vs. novelty and reward information $p=0.276$; permutation tests). Furthermore, neural coding indexes for sensory surprise and recency were not correlated with the analogous indexes for reward value and reward information anticipation at the population level (Figure 3.5C).

It is worth noting that these results do not shed light on the interaction between reward and novelty when the value of novel objects changes across trials, or when novelty and reward vary within the same task, or for example when novelty signals the chance to learn the value of objects (Kakade and Dayan, 2002; Costa et al., 2019; Costa and Averbeck, 2020). Rather, the analyses were designed to fortify the notion that the population-wide results (Figure 3.3A-C) were not simply due to task-induced fluctuations object salience or arousal.

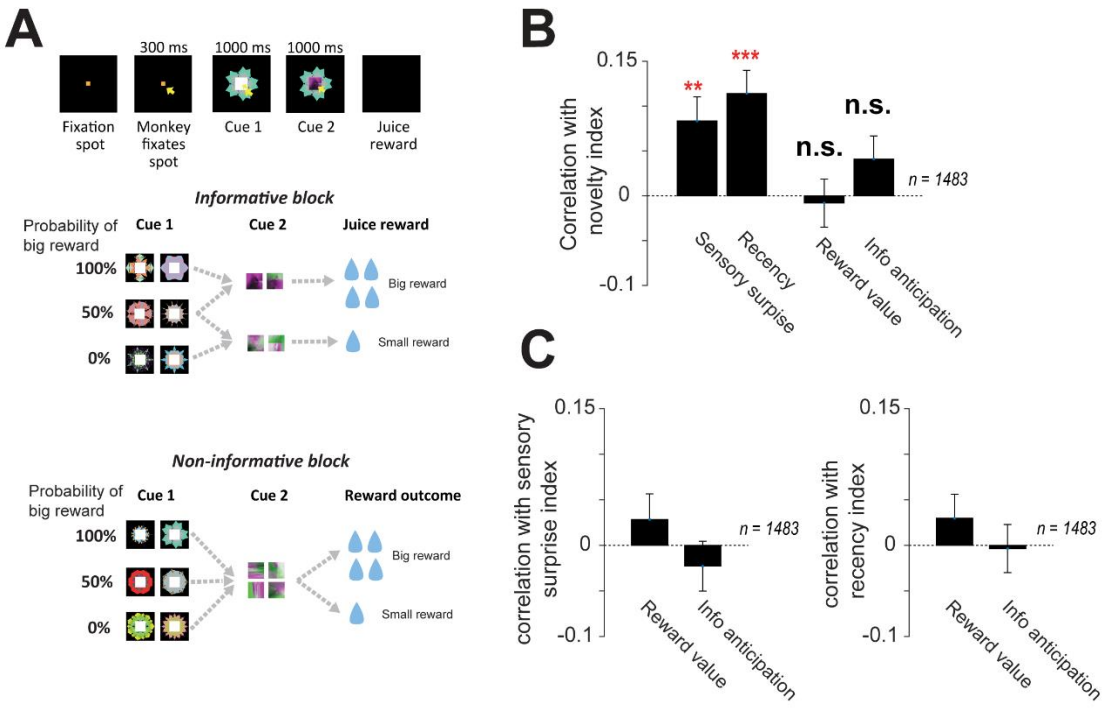


Figure 3.5. Neurons' excitatory responses to novelty in the object viewing procedure are on average not correlated with their responses in a distinct reward information viewing procedure. (**A**) Behavioral procedure for testing neural responses to changes in reward value and in anticipation of information to resolve reward uncertainty. (**B**) Novelty-excited neurons' novelty sensitivity was not significantly correlated with their sensitivity to reward value and information anticipation (right two bars, rho = 0.084, p = 0.0013, rho = 0.114, p<0.0001). These correlations are weaker than the correlation with sensitivity to sensory surprise and recency (left two bars, rho = -0.0008, p = 0.76, rho = 0.042, p = 0.11). *, **, ***, indicate p < 0.05, 0.01, 0.001. (**C**) Novelty-excited neurons' sensitivity to sensory surprise and recency was not significantly correlated with their sensitivity to reward value and information anticipation. (from left to right: rho = 0.0284, p = 0.27, rho = -0.023, p = 0.38, rho = 0.030, p = 0.25, rho = -0.0036, p = 0.89)

Compared to the more prevalent novelty excited neurons, on average, novelty inhibited neurons behaved slightly differently; they had marginally significant correlations with reward related indices but not with sensory surprise or recency indices (Supplemental Figure 3.4C), suggesting that they may have a different role in linking novelty, arousal, and reward processing.

Next we asked whether novelty responses share common origins across simultaneously recorded neurons. We found that during neural responses to objects, noise correlations are higher between pairs of novelty responsive neurons than pairs of other neurons ($p=0.002$, paired signed-rank test; Supplemental Figure 3.5). Furthermore, ensembles of these novelty responsive neurons had significantly expanded variance in the strength of their novelty signals across presentations of different individual novel objects, consistent with the idea that neural systems for novelty detection can have shared response variance, effectively treating some novel objects as 'more novel' and others as 'less novel' (Meyer and Rust, 2018; Mehrpour et al., 2021) (Supplemental Figure 3.5).

3.3.4 Multiple timescales of learning and forgetting exists across neurons and across brain areas

Having delineated key factors underlying novelty processing in the brain, we next set out to measure their timescales of operation. In everyday life, object novelty is fundamentally linked to a continuous process of learning, as each new novel object gradually becomes familiar with repeated experience. Furthermore, this learning can occur at multiple timescales; sometimes rapidly, sometimes slowly, and sometimes interrupted by forgetting. To investigate the timescales of this novelty-familiarity learning, our behavioral procedure (Figure 3.1) included Type 4 trials in which sequences of three objects could contain novel objects that sometimes

repeated across experimental sessions across multiple days, for up to 5 days (Materials and Methods). As a result, these “repeating novel” objects underwent a novelty-to-familiarity transformation, allowing us to measure each neuron's responses during different stages of learning.

We found that the novelty excited neuron population's average novelty-related activity reflected a gradual learning process, marked by repeated cycles of rapid within-day learning in each session, followed by substantial between session forgetting, in between sessions (Figure 3.6). To quantify this across the population of novelty-excited neurons controlling for novelty unrelated sensory adaptation, we normalized neural responses to repeating novel objects based on the responses to fully novel and fully familiar objects in the same trial type and during similar epochs of the behavioral session (Materials and Methods).

We found that on Day 1 of exposure to a repeating novel object, the population response rapidly learned to differentiate them from truly novel objects even after a few exposures (Figure 3.6A). That is, as a novel object repeated within a day, the average novelty related response of novelty neurons declined. Furthermore, there was also a rapid decline in the learning rate – the fraction of novelty-related activity that disappeared with each exposure to the object. To quantify this, we calculated a measure of the neural learning rate from each individual object exposure. This analysis revealed that at the population level the neural learning rate started high and then rapidly declined over the course of the day (Figure 3.6B). Importantly, this was found despite controlling for any potentially confounding impact of arousal or task engagement on the time course of the novelty-familiarity transformation (Materials and Methods).

Despite its rapid within-session leaning, these neurons did not retain the results of their learning perfectly across sessions. Instead, they showed clear evidence of substantial between-session forgetting (Figure 3.6C). When the same neurons were presented with repeating novel fractals that had been experienced for 2+ days, their initial response was substantially lower than their response to a completely new fractal, indicating some retention of learning (Day 2+ start < Day 1 start, $p<0.0001$, paired signed-rank test). However, it was also substantially higher than their response to a repeating novel fractal at the end of the first day, indicating substantial forgetting or loss of access to the prior learning (Day 2+ start > Day 1 end, $p = 0.0099$, paired signed-rank test). Note that this is distinct from the pattern one would expect if neural memories of objects are enhanced following overnight rest between training sessions (Stickgold, 2005). If that was the case, rest should cause neurons to treat the repeating novel object more similarly to a fully familiar object ("overnight or between-session learning"); instead, neurons treated it more similarly to a fully novel object ("overnight or between-session forgetting"). This resembles the "spontaneous recovery" observed in multiple forms of sensory, motor, and motivational learning (Pavlov, 1960; Rescorla, 2004; Smith et al., 2006; Kording et al., 2007), and could occur due to the passage of time, rest, or other factors. Thus on Day 2+, this neural population had to re-start its learning process from an earlier point on the learning curve.

Thus, the time course of novelty learning in the novelty responsive population followed a saw tooth pattern, marked by cycles of within-day learning followed by partial across-day forgetting (Figure 3.6C). This saw tooth pattern continued with striking consistency on each day for up to the full 5 days of exposure tested in this experiment, as the population response to the repeating novel fractals gradually progressed toward familiarity (Figure 3.6D). The population average response to repeating novel fractals never became fully identical to familiar fractals (Figure

3.6C, $p<0.0001$, Figure 3.5D, $p<0.0001$, signed-rank test relative to 0.5), which is expected since the familiar fractals had been previously viewed many times (1000+ exposures). Interestingly, the population made quite similar learning progress each day; the learning rate had a remarkably similar magnitude and time course over the session for fractals regardless of their number of days of exposure (Figure 3.6B, Day 1 vs. Day 2).

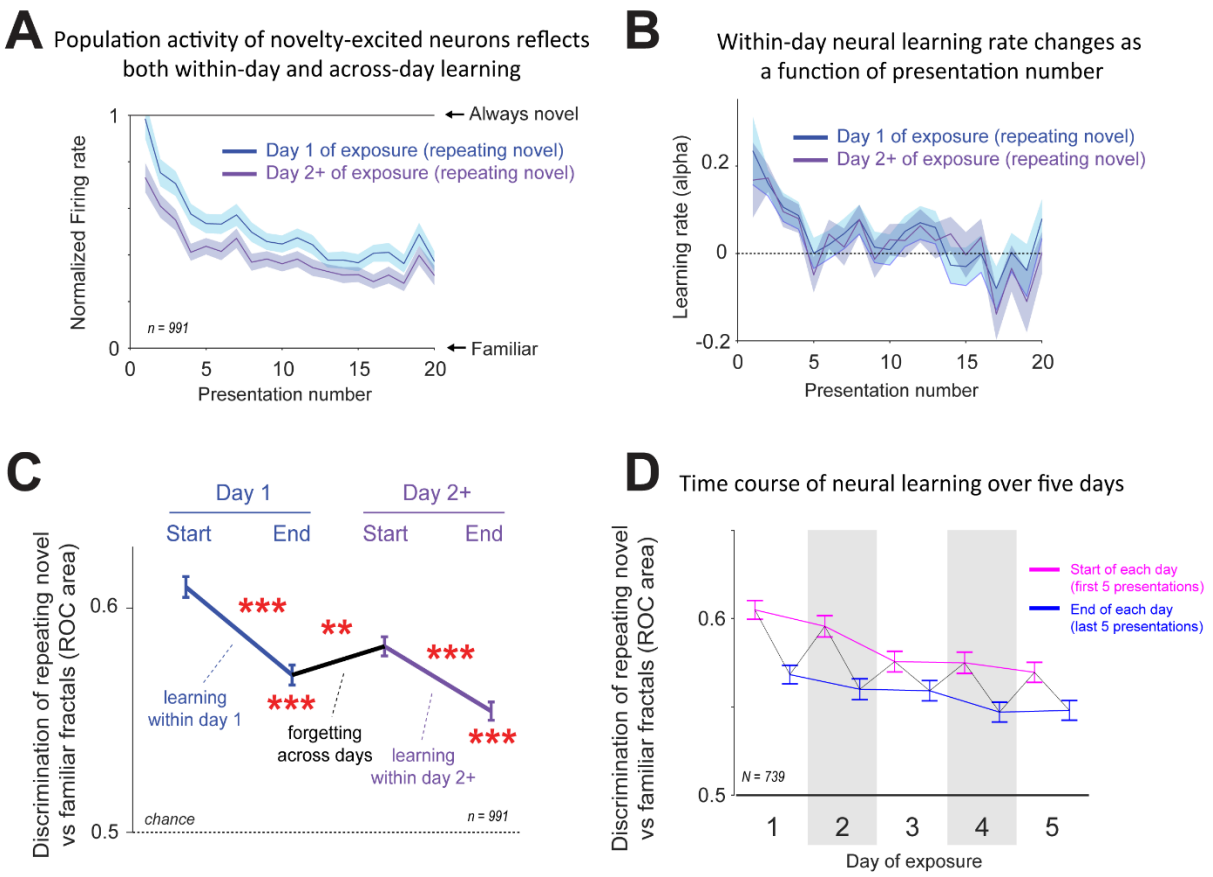


Figure 3.6. Dynamics of learning and forgetting in novelty excited neurons. (**A**) Novelty-excited neuron population response as a function of object exposure. Firing rates are normalized here so that the response is 0 to familiar objects and 1 to always novel objects (Materials and Methods). (**B**) Novelty-excited neuron population learning rate decreases over the course of each day. (**C**) Quantification of within-day and across-day novelty-familiarity transformations. The y axis is the population mean AUC of the ROC for discriminating responses to repeating novel fractals vs. familiar fractals. This is quantified for the first 5 (start) and last 5 (end) presentations of each repeating novel object on each day. *, **, *** indicates $p < 0.05$, 0.01, 0.001 (signed-rank tests). (**D**) Novelty-excited neurons across day novelty-familiarity transformations over the course of 5 days (Monkey S; Materials and Methods). The y axis is the area under the ROC curve comparing responses to repeating novel fractals vs. familiar fractals. Shown is data from the first 5 (**magenta**) and last 5 (**blue**) presentations of each fractal in each session.

So far, we found that novelty-excited neurons as a population can have different learning rates at different times within a single session. But could there also be variation in learning across the population, such that different neurons learn and forget at different rates? Some of the novel objects we encounter in life are only relevant to our immediate situation, but others have long-term importance and must be committed to memory, sometimes forever (Hikosaka et al., 2013). Therefore, it would be ideal if the brain contained neural networks with different learning rates to handle these diverse situations. Indeed, it has been proposed that the brain contains reservoirs of neurons with different timescales of learning, including 'fast' neural systems that both learn and forget quickly, and 'slow' systems that both learn and forget slowly (Smith et al., 2006; Kording et al., 2007). Alternately, it is possible that novelty responsive neurons throughout the brain learn and forget in lock step with each other, cooperating to form a single, unified representation of each object's degree of novelty. This would be analogous to theories of motivated behavior, which propose that key motivational variables, such as the values of states and actions, are computed once and then used by multiple brain areas to guide multiple processes such as learning, outcome anticipation, and decision making (Schultz et al., 1997; Schultz, 2002; Padoa-Schioppa and Cai, 2011). While each of these hypotheses from the literature on learning and memory has large implications for brain function, they have remained untested in the realm of novelty detection.

To answer this open question, we examined novelty responsive neurons and found that they learned in diverse manners. For example, Figure 3.7A left panel shows an example novelty responsive neuron recorded in area AVMTC, suggestive of 'slow learning, slow forgetting'. That is, this neuron was strongly excited by its first exposure to repeating novel objects (Day 1 start), gradually reduced its response over the course of the day (Day 1 end), had almost the same level

of response at start of 2+ day, which showed roughly no forgetting (Day 2+ start), and produced a learning curve that was similar to the first day but with a pronounced downward shift, suggestive of progressive learning across days (Day 2+ start to end). By contrast, Figure 3.7A right panel shows a second AVMTC novelty responsive neuron with very different learning curves, suggestive of 'fast learning, fast forgetting'. This neuron learned at a considerably faster rate on Day 1: by the end of the day its response to repeating novel objects is very close to fully familiar objects. However, this neuron also had considerably greater rest related forgetting: it almost completely 'reset' its novelty response on Day 2+, so that the learning curve on Day 2+ has only a little downward shift from Day 1. Thus, while this neuron learned rapidly within each day, it was unable to retain this learning and hence could not compound its progress across days.

To quantify the time course of novelty-familiarity transformations in single neurons, we calculated independent indexes of how much a neuron learned within a day and forgot across days (Figure 3.6C, Materials and Methods) and assessed the relationship between learning and forgetting. We found that the two indices were strongly correlated (Figure 3.7B, $\rho=0.371$, $p<0.0001$, Spearman's rank correlation). This correlation was driven by true variation across neurons in their learning-related activity, not by variations across session in the animal's learning process, because it remained similarly strong after subtracting out session-level effects (Supplemental Figure 3.6B). These results indicate that there was true variation in the degree of neuronal learning, such that neurons that learned a greater amount within Day 1 also tended to forget a greater amount across days. Moreover, while within brain regions these types of neurons were intermingled (e.g. the cells in Figure 3.7A), we did observe an anatomical trend that strongly corroborates the results of Figure 3.6C: brain areas that on average had greater within day learning also had greater across day forgetting (Figure 3.7C). This result supports theories

that anatomical or circuit differences can support learning on different time scales (Bromberg-Martin et al., 2010b; Kim and Hikosaka, 2013; Murray et al., 2014; Cavanagh et al., 2016; Monosov et al., 2020; Spitmaan et al., 2020), and provides evidence that this is the case specifically in the realm of novelty-to-familiarity transformations.

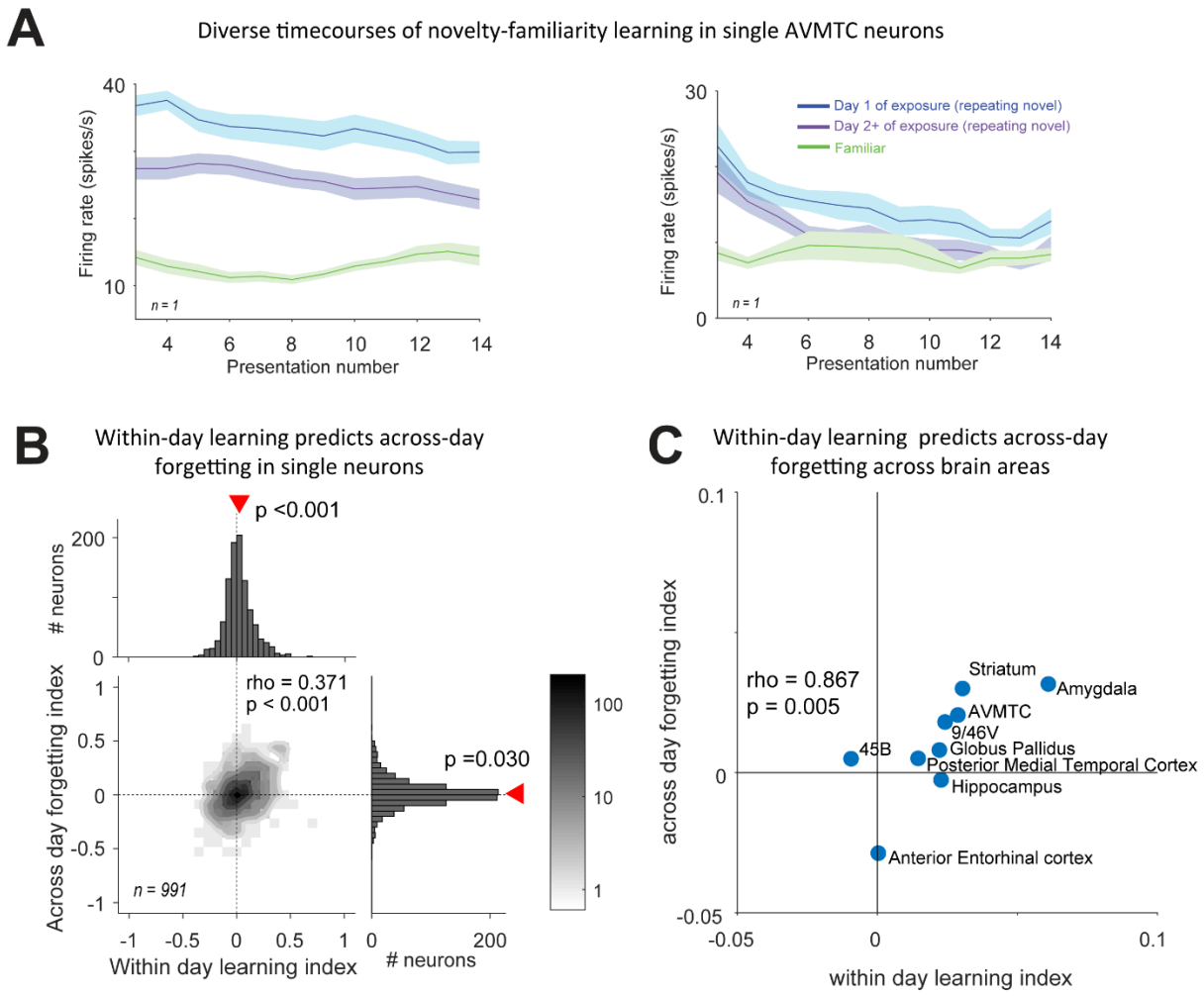


Figure 3.7. Multiple rates of learning and forgetting across neurons and across brain areas. (**A**) (**left**) Activity of an example AVMTC neuron with slow, progressive learning over the course of multiple sessions. This plot was made by quantifying activity using a sliding window of 5 object presentations, advanced in steps of 1 object presentation; error bars indicate SE. (**right**) A second example AVMTC neuron that learned rapidly within each day but almost completely reset its learning curve across days. (**C**) Correlation between indexes of neural learning within day (x-axis) and forgetting across days (y-axis). The heat map shows the joint distribution of the two indexes, and the histograms show the marginal distributions of each index. A positive within-day learning index indicates responses to repeating novel fractals become relatively more like responses to familiar fractals at Day 1 end than Day 1 start. A positive across-day forgetting index is an analogous comparison between Day 1 end and Day 2 start. (**D**) Heterogeneity in average learning and forgetting indices across brain areas. Scatterplot showing the distribution of the mean within day learning index and mean across day forgetting index of each brain area.

3.4 Discussions

Current theories put forward distinct models in which novelty is either derived from recency judgements or conceptualized as a form of sensory surprise. Here, high-channel electrophysiology assessed these mechanisms and determined that object-novelty arises with both computations of recency and surprise, suggesting that novelty detection and memory-related functions are supported by diverse mechanisms of predictive coding.

Elegant work from Brown and his colleagues identified distinct neuronal groups in AVMTC – the region most strongly enriched with novelty sensitive neurons (Figure 3.4A) – that signaled novelty and recency (Xiang and Brown, 1998). There the relationship between recency and novelty, or surprise and novelty, was not assessed. Subsequent pioneering theoretical and computational work of Bogacz and Brown suggested that recency and novelty neurons may form a functional network in support of memory and other adaptive behaviors (Bogacz et al., 2001a; Bogacz and Brown, 2003b; Bogacz and Brown, 2003a). Our data replicate the findings that AVMTC is enriched with novelty sensitive neurons, above and beyond other temporal regions, and provides a first demonstration for the linkage between recency and novelty computations suggested by modelling. We could have found instead that novelty sensitivity was not correlated or negative correlated to recency or surprise. In some sense, within a single pool of neurons such as AVMTC, this could have been more efficient mode of information encoding (Koay et al., 2021). This deviation from efficiency further supports the argument that the relationships between novelty and sensory surprise and recency are important for novelty detection.

Importantly, our data does not indicate that novelty, recency, and sensory surprise are always treated identically by the brain (Bogacz et al., 2001a; Charles et al., 2004). Instead, we interpret

the results to mean that neurons sensitive to these variables are functionally linked in support of novelty detection in the primate brain, across multiple brain areas (Bogacz et al., 2001a). Novelty triggers many processes, starting with the retina (Hosoya et al., 2005; Huang and Rao, 2011), in the service of adaptive behavior and survival. Next, in-vivo and computational experiments must understand how bottom-up novelty-related computations and top-down (feedback) signals interact to produce the complex interactions between different prediction signals, and determine which emerge due to bottom-up sensory processing.

For example, novelty and surprise can arise due to distinct sources. Sensory surprise sensitivity indices (Figure 3.1) measure surprise due to probability of seeing an object, while surprises due to sequence violations (Supplemental Figure 3.3B-D) arise *also* due to a violation in the subjects' beliefs about the structure and statistics of the external world. Relatedly, novelty can be expected or unexpected, arising from distinct sources ((Barto et al., 2013; Zhang et al., 2019; Monosov, 2020), note that in our data sensitivities to unexpected novelty in Type 4 trials and to expected novelty in Type 1 trials were highly correlated; $\rho=0.3582$; $p<0.001$). While expected and unexpected novelty or surprises formally could have different roles in learning (Soltani and Izquierdo, 2019), in naturalistic contexts, the line between expectedness and unexpectedness is never clear – agents need to monitor all surprising and novel events to detect changes in contexts and distributions of outcomes and events (Monosov, 2020). This notion is supported by the linkage we see between novelty sensitivity and different forms of sensory surprise across many brain areas (Figure 3.3-4, and Supplemental Figure 3.3D). However, the source of sensory surprises along different task, decision, and statistical *hierarchies* requires further investigation as has begun to be done for reward surprises or prediction errors (Li et al., 2019; Sarafyazd and Jazayeri, 2019).

Beyond the global linkage between novelty with sensory surprise and recency, we found several brain area differences (Figure 3.4). While, the same neurons within ventral visual regions in AVMTC were strongly sensitive to both sensory surprise and novelty, single novelty sensitive neurons there did not seem to commonly track object recency relative to other areas. In particular, novelty sensitive neurons in the amygdala (Figures 3.2 and 3.4) were particularly also sensitive to sensory surprise, recency, and novelty consistent with the wide ranging roles of amygdala neurons in object memory, sensory processing, and associative learning (Murray and Mishkin, 1998; Baxter and Murray, 2002; Murray and Izquierdo, 2007; Peck et al., 2013; Peck and Salzman, 2014; Dal Monte et al., 2015; Costa et al., 2016).

Neuronal novelty-familiarity transformations occurred with multiple, heterogeneous time courses, with some neurons learning slowly but steadily over the course of multiple days, and others learning a rapidly within each day but retaining little across days. This may support adaptive behavior in a world in which some objects are only relevant for short periods of time, while others must be remembered for a lifetime. This is akin to theories of sensorimotor learning and adaptation in which the motor system must adjust to perturbations during movements that are unlikely to ever again challenge the system, and also to slow permanent changes, like age-related changes in the body or permanent changes in the task (Smith et al., 2006). It has been suggested that to support such fast and slow changes the brain may have evolved fast and slow learning systems (Logothetis and Sheinberg, 1996; Kording et al., 2007; Kim and Hikosaka, 2013), that forget quickly or slowly, respectively. We found evidence for similar mechanisms for novelty detection. Neurons underwent novelty-to-familiarity transformations at a spectrum of timescales, with some learning rapidly and others learning slowly. Furthermore, these timescales

of learning were reflected in a consistent manner across time; neurons which learned rapidly within a session also tended to have greater spontaneous recovery across sessions.

We found that the entorhinal cortex and the hippocampus had average negative forgetting indices, meaning that their average learning was enhanced *across* days, after periods of rest (Figure 3.7 and Supplemental Figure 3.6C). While among single neurons, all brain regions contained heterogeneous timescales, consistent with the average learning and forgetting indices, posterior medial temporal cortex, anterior entorhinal, and hippocampal regions were also enriched with single neurons that displayed enhanced learning after rest (Supplemental Figure 3.6C). We propose that this enhancement (indexed by negative forgetting indices) is a correlate of these regions' pronounced contributions to memory consolidation and recall (Saunders et al., 1984; Murray and Wise, 1996; Murray et al., 1998; Hasselmo and McClelland, 1999; Hasselmo et al., 2000; Fell et al., 2002; Joo and Frank, 2018; Pine et al., 2021), but also caution that many brain areas contained heterogeneous learning and forgetting timescales, regardless of the average timescales (e.g., negative or positive). This is consistent with the notion that learning and memory are distributed functions rather than pinpointed to any one particular region of the brain.

Neural learning-forgetting timescales could be directly related to the computations of recency and surprise. Intuitively, the faster we forget, the more surprising the forgotten object becomes. Despite the fact that our task painstakingly separated recency and surprise, their relationship with novelty suggests that timescales may play a central role in novelty detection, albeit in heterogeneous manners, and that novelty detection is a dynamic process that may not be bound to binary classifications of novelty and familiarity.

A technological limitation in our study was that while we were able to record from many regions simultaneously, each electrode had one contact. Therefore, each recording session yielded relatively few neurons within each region. Denser sampling within each brain area could facilitate hierarchical analyses of learning and forgetting timescales across brain areas, and relate them to intrinsic timescales (Murray et al., 2014; Cavanagh et al., 2016; Spitmaan et al., 2020) or anatomical wiring. Also, to perform high dimensional population analyses, Materials and Methods for wide-scale recording that incorporate denser sampling of neurons within each brain area are required.

Surprises may arise due to different circuit and algorithmic mechanisms in the service of distinct strategies of prediction (e.g., depending on the coding space or reference frame of the prediction or prior). This is well illustrated by the insight gained in the inferotemporal cortex (IT), lateral to AVMTC. Kaliukhovich and Vogels (2014) found that IT neurons responded more strongly to an object in blocks when it had a 10% probability of appearing rather than a 90% probability of appearing, and this adapted rapidly to recent stimulus history (Kaliukhovich and Vogels, 2014). These responses were strikingly insensitive to the prior probability distribution of alternative objects that *could* have been presented, to either a narrow distribution (a single alternative object with a 90% probability of appearing) or to a wide one (9 different alternative objects which each had a 10% probability of appearing). Meyer and Olson (2011) trained monkeys with highly stable sequences of objects (i.e., object A->B, A->B, C->D, C->D) and after extensive training violated the monkeys' expectations (i.e., C->B). IT activations were higher in response to these sequence violation events which they attributed primarily to the suppression of predicted stimuli (and to the absence of this suppression for objects that violate predictions (Meyer and Olson, 2011; Ramachandran et al., 2016). A study by Bell et al (2016) reported that IT responses to

faces were suppressed during periods when faces were presented with high probability (Vogels, 2016; Bell et al., 2017). In sum these papers suggest that surprise effects in IT could be due to computations tightly linked to object-presentation probability.

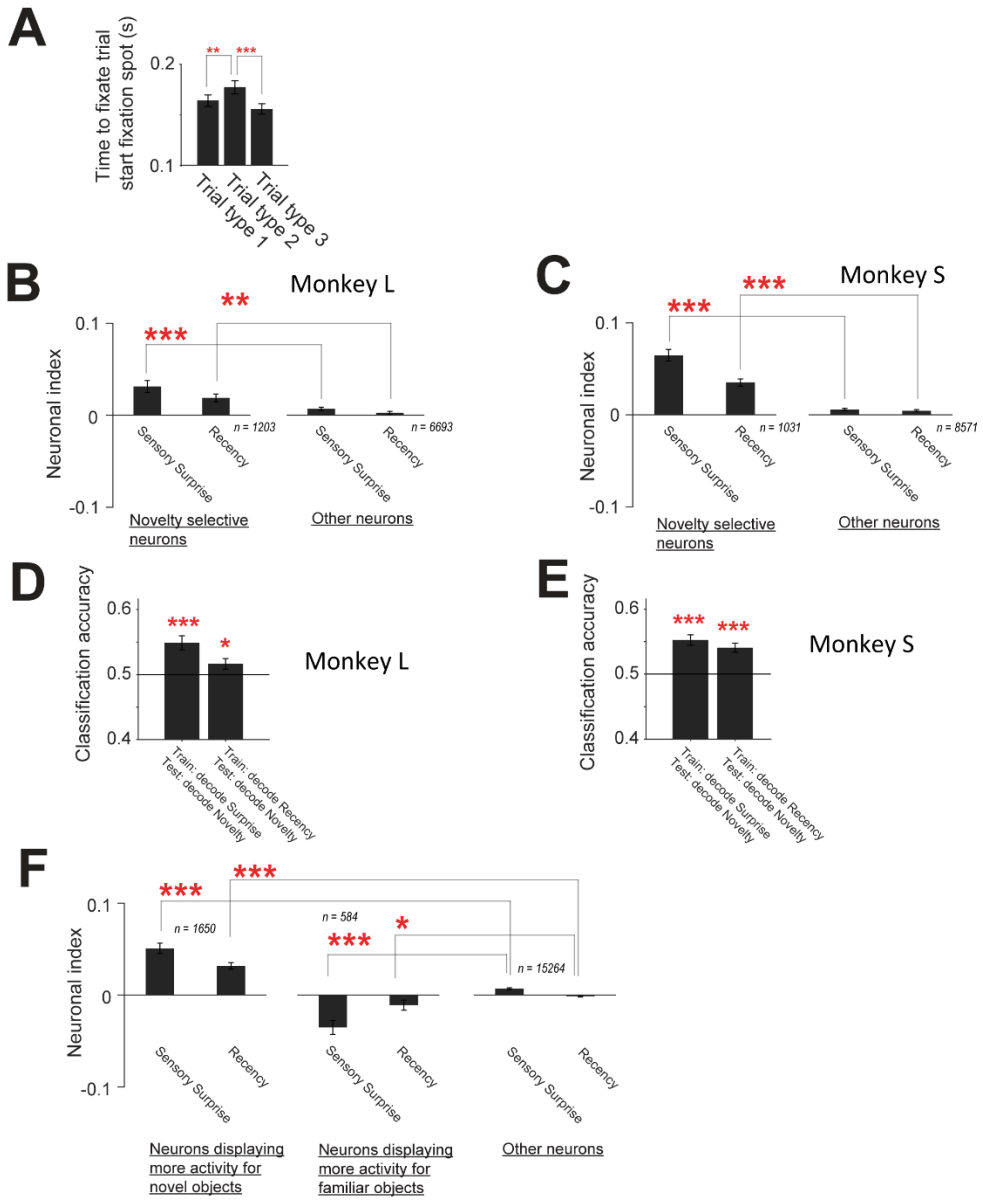
Our data in the AVMTC, constituting medial regions of IT and perirhinal cortex (Xiang and Brown, 1998), is aligned with this interpretation. When comparing activity in trial types 2 and 3, which differ in monkeys' ability to predict which object will appear next (object-presentation probability), we observed differential activation in many AVMTC neurons, particularly in neurons that also displayed strong selective activations in response to novel objects. Here, we show that novelty selectivity - the differential responses to never before seen objects - could arise from computations that detect the occurrence of unpredicted objects (or surprises (Barto et al., 2013)), emerging intertwined with the circuit mechanisms of prediction (Aitchison and Lengyel, 2017).

It is highly possible that surprise signals in different brain regions serve different prediction and behavioral control algorithms, supporting different forms of belief or sensory updating. Such diversity can theoretically support the many distinct strategies of novelty detection that have been developed in machine learning (Miljković, 2010). For example, different brain areas or circuits may implement distinct algorithms based on comparisons of explicit sensory memories with incoming sensory stimuli (Dasgupta et al., 2018; Tyulmankov et al., 2021), such as in Hopfield networks (Bogacz et al., 2001b; Bogacz and Brown, 2003b), or implement algorithms for inference, comparing beliefs about a particular object's probability, or total contextual object variability, with sensory input. Diversity of novelty detection and memory encoding mechanisms

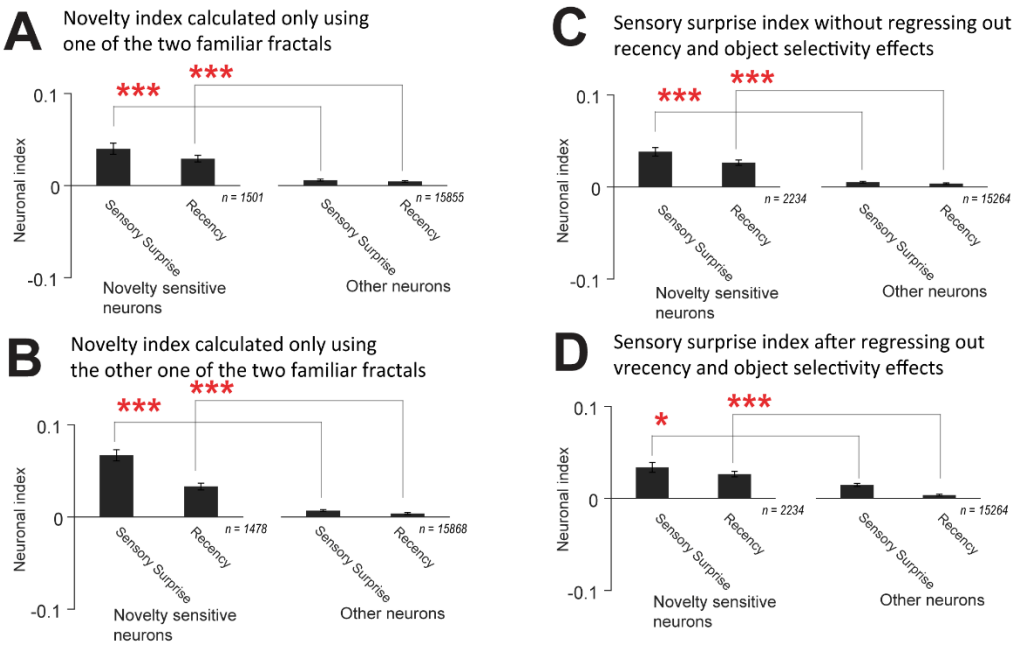
may be supported by the heterogeneous time scales of learning and forgetting that we observed within individual brain regions, and across the brain.

Therefore, a particularly fruitful avenue for future work will be to uncover how multiple timescales of learning and forgetting can sub serve adaptive novelty-detection and inference in natural and artificial intelligence (Adams et al., 2014; Aitchison and Lengyel, 2017; Friston, 2018).

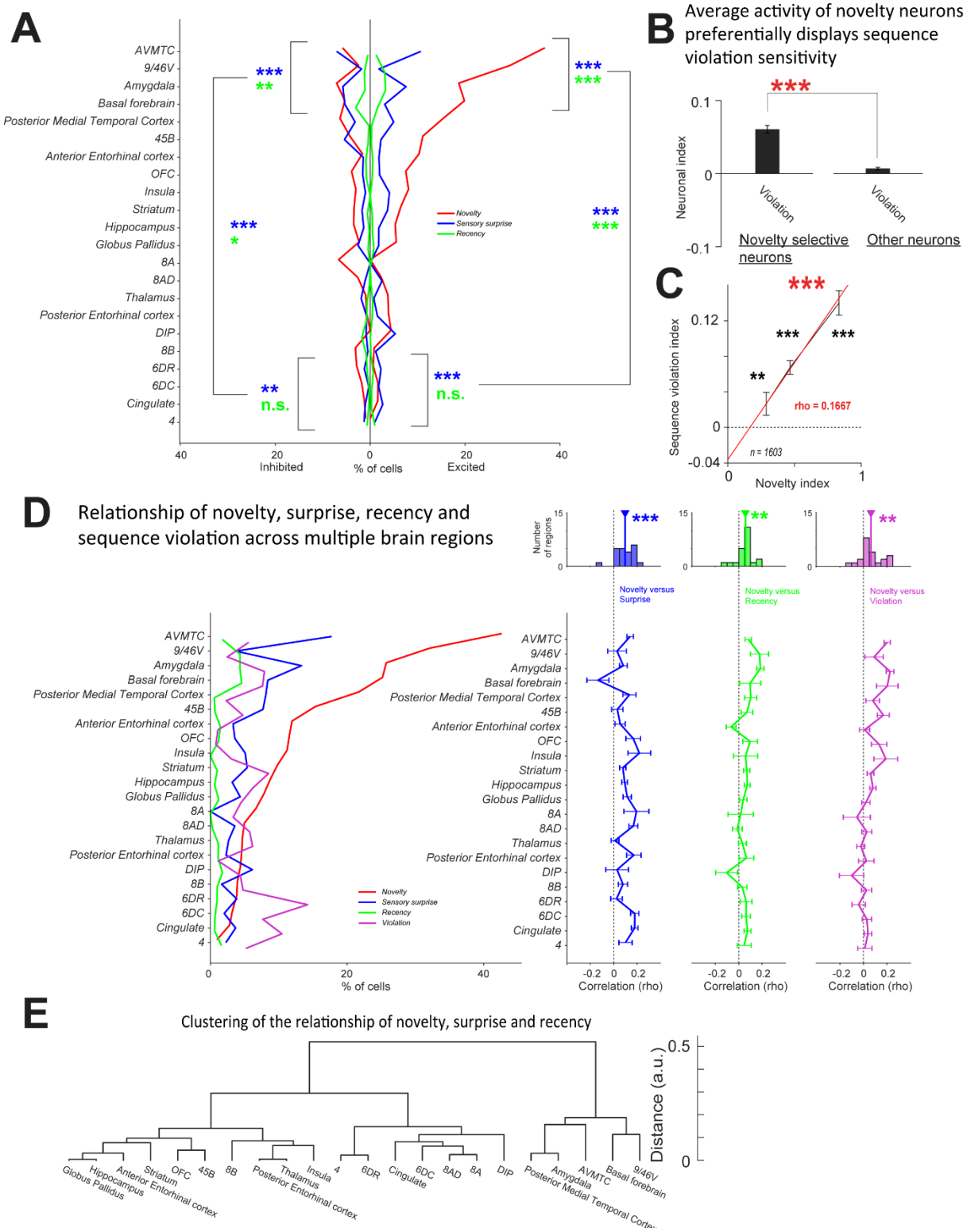
3.5 Supplemental materials



Supplemental figure 3.1. Trial start related behavioral analyses and neural results for two monkeys separately. (**A**) Monkeys' reaction times are different for the fixation dots predicting the first three trial types. (**B-E**) In both monkeys, novelty neurons displayed strong coding of sensory surprise and recency. The figure format is the same as in Figure 3. (**F**) Novelty neurons in Figure 3A shown separately for novelty-excited (higher activity for novel objects - left two bars) and novelty-inhibited (higher activity for familiar objects - middle two bars) groups. Other neurons are shown on the right (same as Figure 3). The figure format is the same as in Figure 3.

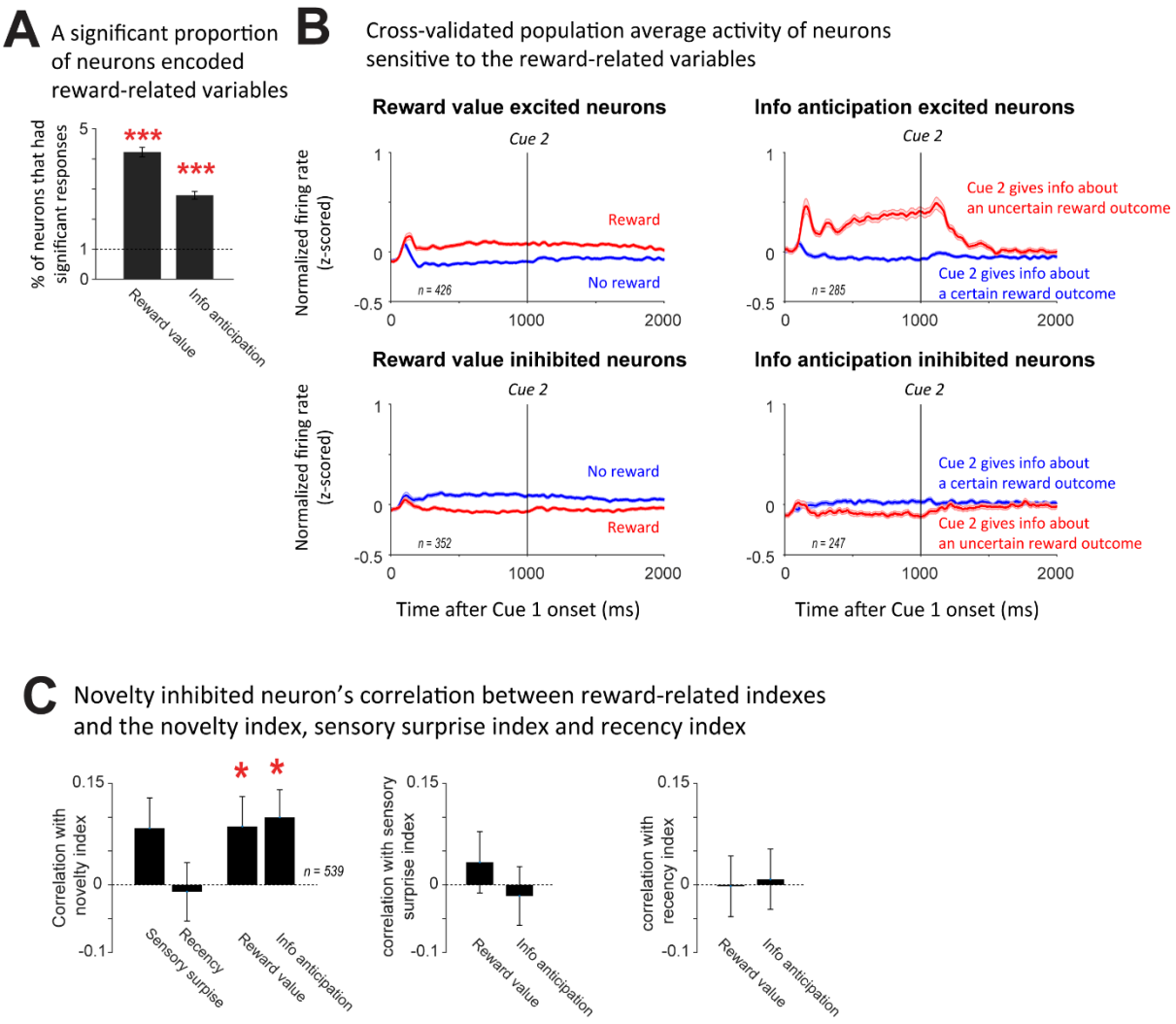


Supplemental figure 3.2. Supplemental analyses for novelty and surprise indices. Correlation analysis as in Figure 3A, but using alternative Materials and Methods to calculate the novelty index and sensory surprise index as controls. (**A**) We calculated the novelty index only using one of the two familiar fractals in trial type 2 at the second position in the sequence. The sensory surprise index and recency index are very similar to Figure 3. (**B**) We calculated the novelty index only using the other one of the two familiar fractals in trial type 2 at the second position in the sequence. The sensory surprise index and recency index are again very similar to Figure 3. (**C**) We calculated the sensory surprise index using the raw neural firing rates without regressing out recency and object selectivity effects. (**D**) We calculated the sensory surprise index after regressing out recency and object selectivity effects from each neuron's firing rates (Materials and Methods). In addition, the correlation coefficient of the novelty index and sensory surprise index in novelty excited neurons is 0.078 ($p<0.01$), 0.071 ($p<0.05$), 0.108 ($p<0.001$), 0.064 ($p<0.01$), respectively in the four cases, and the correlation coefficient of the novelty index and recency index in novelty excited neurons is 0.117 ($p<0.001$), 0.038 ($p=0.20$), 0.114 ($p<0.001$), 0.114 ($p<0.001$), respectively in the four cases.

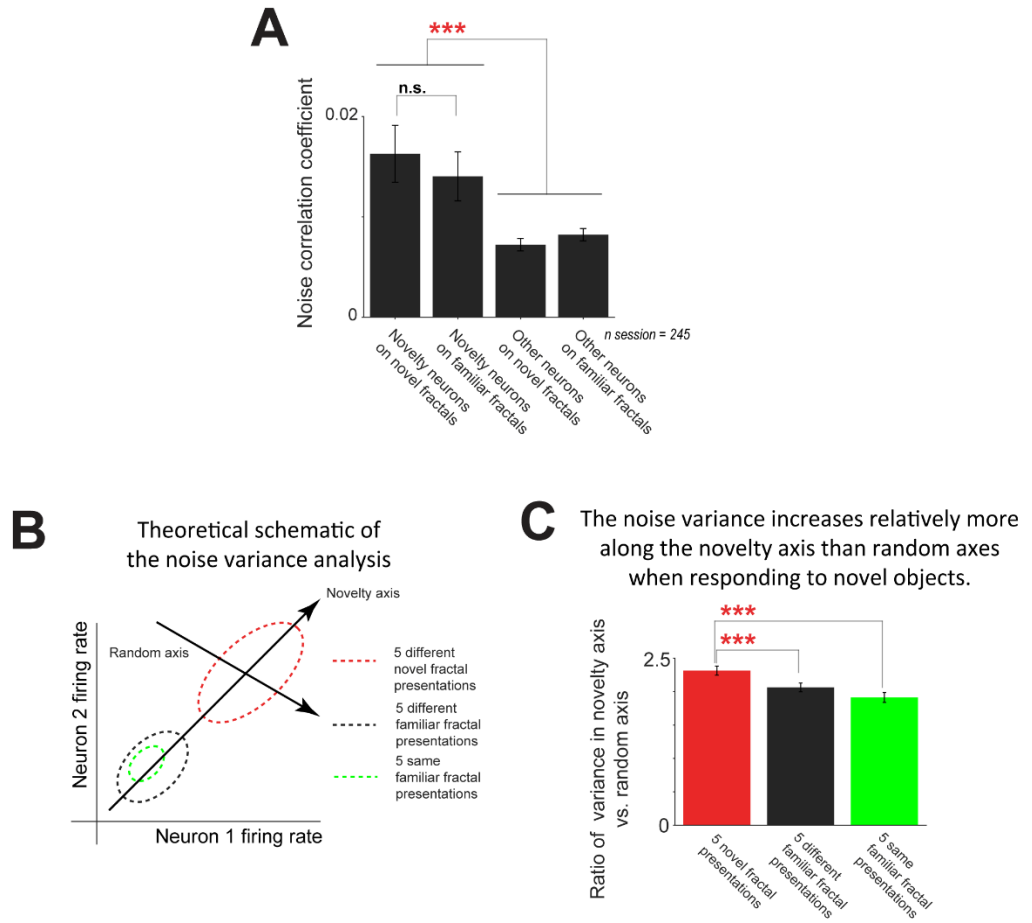


Supplemental figure 3.3. Novelty-excited and inhibited neurons' relationship with surprise and recency across the brain, and sequence violation coding. (**A**) We separated the neurons in figure 3C into novelty-excited, novelty-inhibited, sensory surprise-excited, sensory surprise-

inhibited, recency-excited and recency-inhibited neurons. Data to the left of 0 are for inhibited neurons, and to the right of 0 are for excited neurons. Novelty, sensory surprise, and recency are indicated by red, blue, and green. Here by excited we mean higher activity, and by inhibited we mean lower activity; for example, a novelty inhibited neuron had lower firing rate for novel than familiar objects (Materials and Methods). (**B-D**) Relationship of sequence violation and novelty coding. (**B**) The mean sequence violation index in all novelty responsive neurons is significantly higher than in all other neurons. (**C**) Within novelty-excited neurons, the magnitude of novelty sensitivity was correlated with the magnitude of their sensitivity to sequence violations. The figure format is the same as in Figure 3. (**D**) The distribution of neurons coding sequence violation by brain areas is distinct from Novelty neurons (**left**, magenta line), but in the brain areas which had dense novelty neurons, the coding of novelty and sequence violation are correlated (**right**, magenta). Error bars indicate SE obtained through a bootstrapping procedure. The distribution of the correlations from brain areas is centered higher than 0 (signed-rank test, *, **, ***, indicate $p < 0.05$, 0.01, 0.001). (**E**) Brain areas are clustered by their average recency and sensory surprise sensitivities relative to their average novelty sensitivity (Materials and Methods).

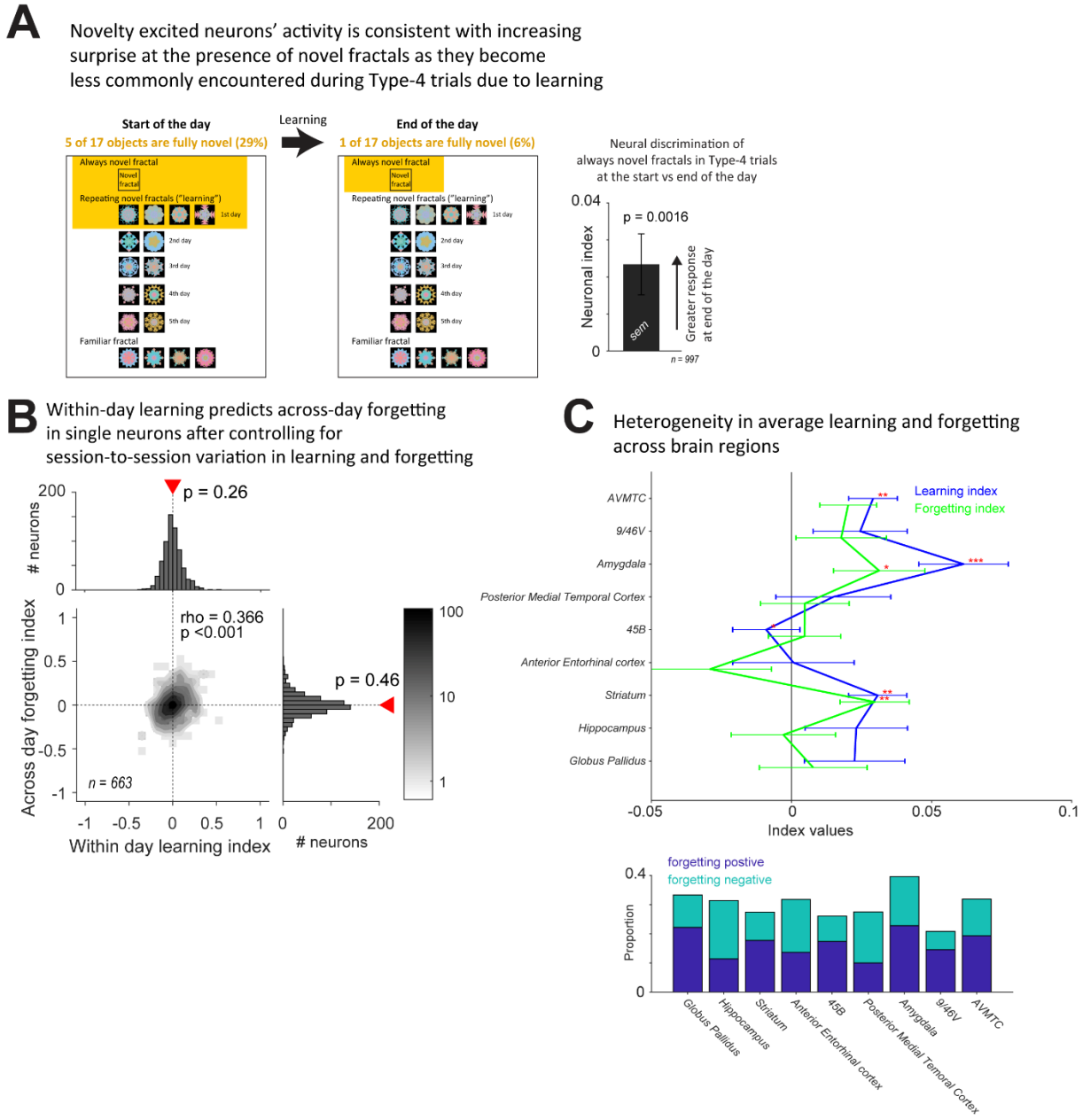


Supplemental figure 3.4. Supplemental analyses of reward information viewing procedure. (**A**) Percentage of neurons significantly encoding reward value and information anticipation. Chance is indicated by the dotted line. (**B**) Cross-validated population average PSTHs of neurons encoding reward value (**left**) or information anticipation (**right**). Each neuron's activity was normalized by z-scoring before being averaged. **(C)** (**left**) Novelty-inhibited neurons' novelty sensitivity correlated with their sensitivity to reward value and information anticipation (right two bars), but not significantly correlate with their sensitivity to sensory surprise and recency (left two bars). *, **, ***, indicate $p < 0.05, 0.01, 0.001$. The results of these comparisons are: novelty and reward value, p=0.045, novelty and information anticipation, p=0.02, novelty and sensory surprise, p=0.052, novelty and recency, p=0.81, Spearman's rank correlation). This implies that novelty-inhibited neurons are functionally distinct from novelty-excited neurons. (**middle, right**) Novelty-inhibited neurons' sensitivity to sensory surprise and recency was not significantly correlated with the magnitude of their sensitivity to reward value and information anticipation.



Supplemental figure 3.5. A hidden factor in the noise correlation of novelty responsive neurons. (**A**) Noise correlation analysis. The bars from left to right represent the mean noise correlations for the following conditions: Novelty responsive neurons' noise correlation during responses to novel fractals; novelty responsive neurons' noise correlation during responses to familiar fractals; other neurons' noise correlation during responses to novel fractals, and other neurons' noise correlation during responses to familiar fractals. Asterisks (same format as other figures) indicate significance of a comparison between novelty responsive neurons vs. other neurons comparing mean noise correlations pooled over all fractals. (**B**) Noise variance analysis. We tested if novelty responsive ensembles responded to the novel fractals in a manner consistent with novel fractals having different degrees of novelty – for example, as a result of some novel fractals being perceived as more or less novel/familiar. Specifically, we defined the ensemble response to each individual object presentation as a point in an N-dimensional firing rate space (with each dimension corresponding to the firing rate of a specific neuron, including only novelty responsive neurons). Then, using 5 individual object presentations, we computed the variance of the ensemble responses along the novelty coding axis and a random axis, bootstrapped the mean, and calculated their ratio (Materials and Methods). This was done for three sets of object presentations: 5 different novel fractals, 5 different familiar fractals, and 5 presentations of the same familiar fractal (red, black, and green circles in theoretical schematic in B; left, middle, and right bars in C). (**C**) The results indicated that the variance of the neural

responses to novel fractals, compared to familiar fractals, was expanded relatively more along the novelty axis than the other random axes ($p<0.0001$, signed-rank tests). This may suggest that neural systems for novelty detection can have shared response variance, effectively treating some novel objects as 'more novel' and others as 'less novel' (Meyer and Rust, 2018; Mehrpour et al., 2021).



Supplemental figure 3.6. The changing of the sparseness of novel objects within a session and supplemental analyses of learning and forgetting. (**A**) In Type-4 trials, as animals learned to become familiar with the repeating novel fractals after repeated exposure, this caused the probability of encountering never-before-seen novel fractals to become much less common (Left panel; start of session: 5/17; end of session: 1/17). Hence, if animals tracked this reduction in the probability of encountering completely novel objects, they might treat the presentation of the always novel objects as increasingly 'surprising'. Furthermore, we hypothesized that novelty responsive neurons might be sensitive to this form of surprise (i.e., surprise induced when animals predict that a novel object has a low probability of appearing), given our finding that many of these neurons respond to a different form of surprise (i.e. sensory surprise induced when

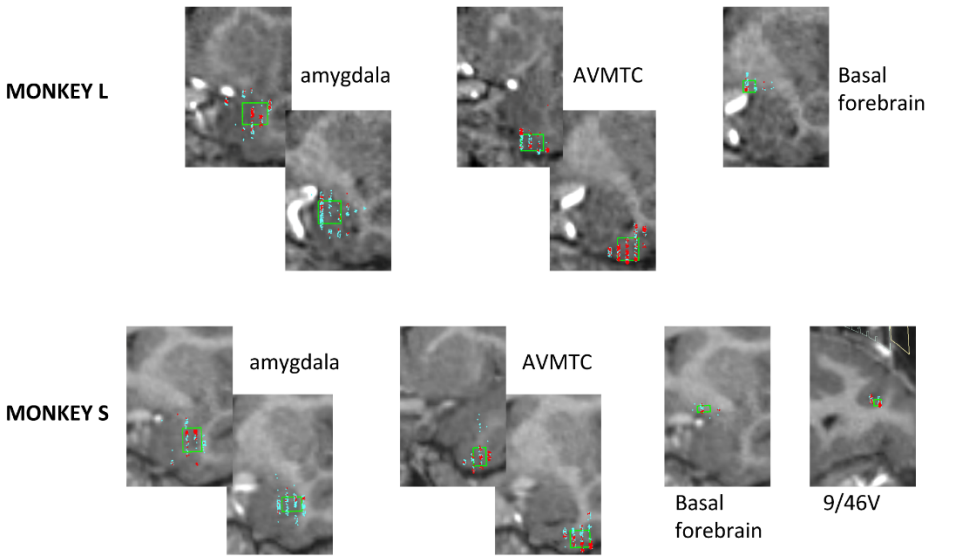
a familiar object has a low probability of appearing; Figure 3). Indeed, the activity of novelty-excited neurons reflected this additional novelty-related surprise by increasing their responses to the always novel objects towards the end of the session. We introduced an index to quantify the effect. Both Type 1 trials and Type 4 trials had novel fractals. In Type 1 trials the percentage of novel fractal was a constant, so we used it to control for any possible drift over time in neural response patterns. The index was calculated as the AUC of the ROC of the neuron's firing rates to the first 5 presentations vs. last 5 presentations of always novel fractals in Type 4 trials minus the AUC of the ROC of the neuron's firing rate to the first 5 presentations vs. last 5 presentations of novel fractals in Type 1 trials. (**B**) Importantly, the correlation between learning and forgetting indexes reflected differences in neural learning, and did not result from any possible session-to-session variations in animal learning or behavior. For example, hypothetically, even if all neurons learned in lock-step with each other within each individual session, if the animal learned fast in some sessions and slow in other sessions, this would produce a dataset where some neurons had fast learning curves and other neurons had slow learning curves. To control for this possibility, we repeated the analysis after subtracting the mean of the indices for each session's data from all neurons recorded during that session (Materials and Methods). The results were very similar to Figure 3.7B. Note that the marginal histograms are no longer significant because the indexes were mean-subtracted within each session, and hence the mean indexes must be equal to 0. (**C-top**) Data of Figure 3.7C with SEM. Solid lines are the means of the within day learning index and across day forgetting index. Error bars indicate standard error of the mean (SEM). (**C-bottom**) Proportions of cells with significant negative or positive forgetting indices (threshold: p=0.05) are indicated below.

A

Monkey L	totalCell	mean(L) to AC	mean(A) to AC	mean(D) to AC	std(L)	std(A)	std(D)
AVMTC	911	13.2	-1.11	-16.8	1.98	1.45	1.84
Basal forebrain	94	6.16	0.0877	-4.63	1.06	0.309	1.03
Amygdala	684	10.8	-0.179	-9.9	2.76	1.85	1.95
Posterior Medial Temporal Cortex	150	14.9	-6.9	-16.4	0.00129	1.86	0.62
45B/8v	110	13.5	4.73	8.19	1.75	0.000844	0.866
Anterior Entorhinal cortex	503	9.4	-3.01	-16.3	1.4	1.36	0.865
Striatum	561	10.3	-2.1	-0.308	3.56	3.64	6.66
Insula	42	16.4	3.45	-0.451	0.00124	0.575	1.01
Hippocampus	833	11.9	-6.4	-11.2	2.59	3.88	2.3
Globus Pallidus	229	6.95	-2.85	-0.87	1.35	1.66	1.58
8A	56	14.9	2.99	9.21	2.34	0.537	1.19
DIP	114	9.42	-16.6	11.2	1.07	0.428	1.02
Posterior Entorhinal cortex	68	8.56	-7.47	-13.5	0.883	0.00074	1.47
8AD	185	11.7	3.92	11.1	1.42	0.763	1.58
Thalamus	373	7.21	-10.8	-0.202	1.41	2.92	2.15
8B	140	9.04	4.73	13.3	0.995	0.000744	0.914
6DR	81	8.82	4.58	17.1	1.5	0.457	1.12
6DC	283	8.26	1.74	14.4	1.45	1.29	2.79
Cingulate	726	5.92	-5.78	14.6	0.532	5.27	2.17
Area 4	2	5.69	-12	19.2	0	0	0

Monkey S	totalCell	mean(L) to AC	mean(A) to AC	mean(D) to AC	std(L)	std(A)	std(D)
AVMTC	532	14.5	-1.19	-17.2	1.66	1.47	1.39
9/46V	163	12.7	10.8	8.77	0.662	0.698	0.551
Basal forebrain	37	7.2	0.0499	-5.19	1.19	0.945	0.574
Amygdala	404	12.5	-0.787	-11.7	1.79	1.32	1.7
Posterior Medial Temporal Cortex	195	15.9	-6.06	-17	0.764	2.55	1.54
45B/8v	332	14	3.11	8.83	1.45	1.31	1.3
OFC	254	11.8	9.15	1.12	2.33	1.1	1.8
Anterior Entorhinal cortex	41	10	-2.15	-17.6	0.724	1.14	0.68
Striatum	944	12	-3.94	-1.88	3.5	5.1	5.11
Insula	57	14.9	-0.223	0.171	0.91	6.81	4.6
Hippocampus	766	13.8	-6.85	-12.2	2.09	2.88	2.01
Globus Pallidus	665	9.09	-3.33	-1.53	1.98	2.13	1.31
8A	34	14.4	-1.11	11.4	0.735	0.438	0.524
Posterior Entorhinal cortex	194	11.7	-7.94	-14.7	0.877	1.43	2.09
8AD	459	11.9	3.69	14.3	1.6	2	0.864
Thalamus	566	7.72	-9.5	2.78	1.31	2.01	1.98
8B	467	7.9	5.92	14.2	1.47	2.09	1.43
6DR	405	7.67	3.47	16.8	1.43	1.48	0.663
6DC	389	9.51	-1.03	15.9	2.61	2.15	1.4
Cingulate	412	5	7.49	13.2	0.984	3.57	1.15
Area 4	306	9.92	-8.67	17	2.94	1.65	1.27

B



Supplemental figure 3.7. Recording locations (**A**) Recording locations relative to the anterior commissure (AC) in each monkey. All units are in mm. Cell count: number of cells in the region.

mean(L) to AC(L) : mean of lateral coordinates of the cells in the region, referenced to AC. mean(A) to AC(A) : mean of anterior coordinates of the cells in the region, referenced to AC. mean(D) to AC(D) : mean of dorsal coordinates of the cells in the region, referenced to AC. std(L) : standard deviation of lateral coordinates of the cells in the region. (**B**) MRI images and recording sites in areas preferentially enriched (top 4) with novelty neurons. The neurons are projected onto coronal MRI slices (top: monkey L and bottom monkey S). Blue dots represent cells which do not selectively respond to novelty and red cells represent cells which do selectively respond to novelty. Green rectangle shows one standard deviation of cells' coordinates around their means in dorsal and lateral directions. Monkey L : AVMTC is shown on two planes at AP +23.4 and AP +19.6, Basal Forebrain is shown on AP + 19.5, Amygdala is shown on AP +22.2 and AP +20.5. Monkey S : AVMTC on AP +24.1 and AP +18.4, 9/46V on AP +30.3, Basal Forebrain on AP +20.1, Amygdala on AP +20.3 and AP + 18.9. AP - anterior posterior axis. Here numbers are relative to the interaural (where the center of the AC is on average between 20+ and 21+ AP).

Chapter 4: Discussions and Summary

4.1 Basal forebrain encodes salient events

Salience, or the behavioral importance of an object, usually reflects its relationship with reward, punishment, uncertainty, surprise, and novelty (Lin and Nicolelis, 2008; Bromberg-Martin et al., 2010a; Ghazizadeh et al., 2016a; Zhu et al., 2018). Accordingly, salience has multiple complex effects on behavior. Salient objects attract attention and orientation, increase arousal, facilitate learning and memory, and affect reinforcement learning and credit assignment (Ohman et al., 2001; Fecteau and Munoz, 2006; Laurent, 2008; Ponzi, 2008; Ghazizadeh et al., 2016a; Zhu et al., 2018; Radulescu et al., 2019; Yu et al., 2021).

In Chapter 2, we study two kinds of neurons in the primate basal forebrain (BF), phasic bursting neurons (BF phasic neurons) and tonic ramping neurons (BF ramping neurons), which may underlie these functions in distinct manners.

On the one hand, the BF phasic neurons respond to reward conditioned stimuli (CS), and their activity scale with reward amount or probability. Unlike some dopaminergic neurons (Schultz et al., 1997), these neurons do not encode quantitative reward prediction error (RPE). Instead, current data show that their activity scales with prediction error only when rewards are delivered. When the reward is unexpectedly omitted, without any external sensory cue, their activity remains largely unchanged. Based on these data and other observations (Lin and Nicolelis, 2008; Avila and Lin, 2014; Hangya et al., 2015), we propose that BF phasic neurons signal surprises triggered by external stimuli. Furthermore, their external surprise signal encompasses other

dimensions. For example, their activity was rapidly increased by novel objects, and violations of beliefs about sensory statistics.

On the other hand, BF ramping neurons' activity predicts the timing of rewarding, novel, and surprising events, and their ramping activity is highly sensitive to the subjects' confidence in event timing. In the experiments, the BF ramping neurons ramped up their activity in response to the CS indicating reward, and their activities were much more enhanced when the reward outcome was uncertain. In the behavior procedure where there were multiple possible reward delivery time points, the BF ramping neurons ramped up to each time point and dropped the activity after, even if there was no external cue. These data show that the BF ramping neurons tracks the time structure of the task and reflects certain internal variables of the subject. The BF ramping neuron could be a good candidate as the source that continuously regulates the subject's internal state, like level of arousal or attention.

The BF contains cholinergic, GABAergic, and glutamatergic projection neurons. Studies in rodents have identified putative GABAergic and cholinergic phasic bursting neurons and non-cholinergic tonic neurons (Lin and Nicolelis, 2008; Avila and Lin, 2014; Hangya et al., 2015). A future direction is to identify which neurotransmitters are released (or co-released) by the BF phasic and tonic neurons in primates.

4.2 Comparison between the basal forebrain phasic neurons and the midbrain dopamine neurons

We can compare the BF phasic neurons with the midbrain dopamine (DA) neurons. DA neurons and BF phasic neurons shared some similarities. They are both in the neuromodulatory systems and have phasic bursting responses to appetitive and aversive events. However, their tunings or

responses to the salient events are not exactly the same, revealing that they might regulate different functions.

A subset of DA neurons encodes reward prediction errors (Schultz et al., 1997). They respond to appetitive and aversive events in the opposite direction (motivational value-coding DA neurons). Another subset of DA neurons signals unsigned reward prediction errors (motivational salience-coding DA neurons) (Matsumoto and Hikosaka, 2009). In contrast, the BF phasic neurons signal incomplete RPEs, i.e., only when the reward is delivered.

Our study demonstrated that BF phasic neurons' salience signal also encompasses novelty and sequence violation. In contrast, some recent studies demonstrated that DA neurons do not encode value-unrelated novelty and sensory surprise (Nour et al., 2018; Ogasawara et al., 2022). In addition, The DA neurons' latency is slightly longer than the BF phasic neurons (DA: ~100ms, BF phasic neuron:20-100ms) (Schultz, 1998; Redgrave and Gurney, 2006; Lin and Nicolelis, 2008; Hangya et al., 2015).

The popular theory of DA neurons is that they participate in the brain's reinforcement algorithm (Schultz et al., 1997; Kakade and Dayan, 2002; Sutton and Barto, 2018). In the credit assignment process, the animal needs to know both how much the value needs to be updated, and which object should be linked to reward or punishment and updated on the value. The animal hypothetically uses the RPE signaled by the motivational value-coding DA neurons for value updating. On the contrary, there are various speculations on the function of the motivational salience-coding DA neurons. They could participate in linking the salient conditioned stimuli to the reward or punishment; Or they could also regulate other effects of motivational salience,

such as attracting attention and orientation, and increasing arousal. (Kume et al., 2005; Redgrave and Gurney, 2006; Laurent, 2008; Bromberg-Martin et al., 2010a)

According to our results in Chapter 2 and previous studies in BF and DA neurons, A preferable speculation is that BF phasic neurons may participate in initiating the orienting and attention, because BF phasic neurons' signal is more rapid and encompasses more general salient events than DA neurons. On the other hand, motivational salience-coding DA neurons can signal complete unsigned RPEs, and they may work with the motivational value-coding DA neurons in the process of credit assignment and learning the association of sensory cues. However, this speculation remains to be tested by further experiments. Furthermore, the serotonergic and noradrenergic neuromodulatory systems also respond to salient events, and how they interact with the BF and dopaminergic systems remains to be understood (Avery and Krichmar, 2017).

4.3 Computations of novelty in the brain

In Chapter 3, we show that the computation of novelty in the primate brain depends on both sensory surprise and recency. This dependency is observed across neurons, and across brain areas.

However, according to our data, different brain areas do not share exactly the same computation of novelty. This supports the hypothesis that there could be multiple systems to detect and encode novel objects, and to regulate different effects driven by novel objects like arousal, attention, orientation, and learning.

Among the brain areas with high concentrations of novelty-responsive neurons, we found that the BF and amygdala are highly enriched with both recency responsive neurons and sensory

surprise responsive neurons. This confirms and adds more knowledge to what we have found about the BF in Chapter 2. Anatomically, the BF and amygdala have reciprocal projections (Mesulam et al., 1983), and previous studies have also found amygdala encodes novelty and surprise (Blackford et al., 2010; Cheung et al., 2019). In addition, the amygdala and the BF are hypothesized to function as a circuit to regulate spatial attention (Peck and Salzman, 2014). Our results so far align with this hypothesis. In contrast, the BF also has wide projections to other neocortex brain areas, but not all brain areas have similar percentage of sensory surprise and recency responsive neurons as the BF. This means that besides the basal forebrain-amygdala circuit, there are probably other circuits that respond to novelty through other mechanisms and regulate other effects of novelty.

The AVMTC has a very high concentration of novelty-responsive neurons and could be a source of novelty signal that is relatively independent from the BF-amygdala circuit. In our data, the AVMTC is highly enriched in sensory surprise coding, which previous studies have also reported: Kaliukhovich and Vogels (2014) and Meyer and Olson (2011) found neurons in the inferior temporal cortex (IT) responding to surprise and sequence. Our data also showed that the AVMTC has a relatively low concentration of recency responsive neurons, which seems not aligned with the study by Xiang and Brown (1998), who found a higher concentration of recency responsive neurons in the temporal cortex. The difference in the results could be due to the different methods used for measuring recency between the two studies: In Xiang and Brown (1998), The nonrecent objects were defined as the first presentations of the objects and recent objects as the subsequent presentations of those objects. Thus, the last time that the animal saw the nonrecent objects was usually 24 hours before. However, in our study, the nonrecent objects had only ~5 mins intervals between presentations, and the recent objects had time intervals of

less than one second. Thus, we measured recency with a much shorter timescale than Xiang and Brown (1998).

We have an experimental discovery that novelty computation is supported or intermingled with the computation of novelty and recency. However, the circuit details underlying these intermingled computations are still missing.

I will speculate some possible ways how the computation of novelty and the computation of surprise may be intermingled. There are primarily three possibilities: The first is that a circuit's primary goal is to compute surprise, and the novelty is calculated as a side effect; the second possibility is that a circuit's primary goal is the compute novelty and the surprise signal is a side effect; the third possibility is that a circuit may compute novelty and surprise together, and the readout of whether the signal is surprise or novelty relies on combining with other information in the downstream brain areas. The same mechanistic relationship may apply for novelty and recency computations as well, and there are even more possible combinations when recency, novelty, and surprise are considered together. Specifically, circuits may be designed to calculate one or two of them, with the others rising as side effects, or all three computations may be totally intermingled. In the following, I will mainly speculate on the models whose primary goal is to calculate novelty. While the other mechanisms are still possible, their discussion is beyond the scope of this dissertation.

Most recently published novelty detection circuit models are designed to discriminate objects seen for the first time from objects that have been seen many times. In some of those models, recency responsive neurons can come out as a side effect but not surprise responsive neurons

(Dasgupta et al., 2018; Tyulmankov et al., 2022). However, we can make improvements on those models to make them also respond to surprise, which fits our experimental findings in Chapter 3.

The principle of novelty detection and surprise detection have a similarity. They both require memory, but they differ in how memory is used. When decoding absolute novelty, the brain essentially compares the current object with ALL objects stored in the memory. In contrast, for surprise, the brain compares the object with only one or a small subset of predicted objects among all objects stored in the brain. To bring surprise computations into the novelty detection circuit, we can assign weights to the objects stored in the novelty detection circuit, and we put more weight on the predicted objects than the other familiar objects. In some perspective, we try to find some way to integrate the predictive coding models into the novelty detection models (Huang and Rao, 2011; Homann et al., 2017).

For example, the fruit fly Bloom filter model used a set of hash functions to map the representation of objects on a bit array (stored in the weights of the KC -> MBON-03 connection) (Bloom, 1970; Dasgupta et al., 2018). A possible modification is to introduce a mechanism that can further decrease the values in the bits that represent the predicted object (i.e., the bits that the object projects to through the hash functions). Thus, the output value of the circuit for the predicted familiar objects will be lower than the unpredicted familiar objects, and the output to the novel objects is still the highest.

We can also make similar modifications in Rafal Bogacz's Hebbian model (Bogacz et al., 2001b), which is a model based on the Hopfield network. Novelty detection is based on the calculation of the Hopfield network's energy (Hopfield, 1982):

$$Energy(X) = \frac{1}{2} XWX^T \tag{4.4}$$

In this equation, X is a row vector representing the input object, and W is a square matrix representing the weights of connections between neurons. Familiar object inputs, on average, have higher energy (Hopfield, 1982), the expectation of the energy ($E[Energy(X)]$) equals 0 for the distribution of novel objects, equals $N/2$ for the distribution of familiar objects, where N is the neuron number (same as the input dimension) in the Hopfield network (Bogacz et al., 2001b).

One way to account for surprise is to add a new term in this energy function:

$$Energy(X;Y) = \frac{1}{2} XWX^T + \frac{1}{2}\alpha * XWY^T = \frac{1}{2} XW(X + \alpha * Y)^T \tag{4.5}$$

Where Y represents a given predicted familiar object and α is a coefficient determining the strength of the surprise response relative to the novelty response. ($0 < \alpha < 1$).

If the input X is same as predicted familiar object Y, the expectation of the energy is

$$E[Energy(Y;Y)] = E[\frac{1}{2}(1+\alpha) * YWY^T] = E[\frac{1}{2}(1+\alpha) * XWX^T] \tag{4.6}$$
$$= (1+\alpha)N/2$$

If the input X is a familiar object but not the predicted object ($X \neq Y$, suppose X and Y are independently drawn and not correlated), the expectation of the energy is

$$E[Energy(X;Y)] = E[\frac{1}{2} XWX^T + \frac{1}{2}\alpha * XWY^T] = E[\frac{1}{2} XWX^T] = N/2 \tag{4.7}$$

In addition, for novelty input Z, the mean of the energy is

$$E[Energy(Z;Y)] = E[\frac{1}{2}ZWZ^T + \frac{1}{2}\alpha * ZWY^T] = 0 \quad (4.8)$$

Equation (4.5) computes the lowest mean of energy; Then is Equation (4.4), which computes the expectation of the energy when the input is unpredicted and familiar; Last is Equation (4.3), which computes expectation of energy when the input is predicted and familiar. Thus, this system can discriminate not only novel vs. familiar objects, but also surprising vs. non-surprising objects. More rigorous analysis of this modified network, which can give the network's capacity, will need to use the signal-to-noise analysis method, similar to what has been done in Bogacz et al. (2001b).

Returning to recency, all novelty detection models with limited capacity, in theory, all have recency response, and the timescales of the recency response depend on the capacity. We observed that in our experiment, some neurons have very short timescale recency responses, which is around the timescale of minutes. However, this timescale is too long to be explained by neurons' refraction period (timescale of milliseconds) (Hodgkin and Huxley 1952) and too short for a large capacity novelty detection network. To give the reader a sense of the scale, here is a rough estimation of the timescale of a network whose size is similar to the primate memory system. There are about 5 billion neurons in the primate brain, with each neuron having, on average, about 1000 synaptic connections (the estimation here is conservative) (Wildenberg et al., 2021). Even if only a small fraction of the neurons participates in memory, say 1%, this still includes 50 million neurons. We take the theoretical circuit model proposed by Rafal Bogacz et al. as an example (Bogacz et al., 2001b; Bogacz and Brown, 2003a), where the capacity of novelty detection is proportional to the number of synaptic connections, and the coefficient is 0.012 given by the model. Then, a primate's brain could theoretically store about $50\ million *$

$1000 * 0.012 = 600\ million$ objects. Supposing a macaque keeps viewing 1 object per second, without sleeping, it would still need about 19 years to fill up the storage.

One possible explanation for the minute-timescale recency is that there are multiple novelty detection networks in the brain, and the sizes of the networks and the synaptic decay rates can vary. For networks with smaller sizes and faster synaptic decay rates, the storage capacities are small, and they generate the minute-timescale recency response we observed.

4.4 Neural learning and forgetting

We investigated how single neurons adapted as novel objects gradually become familiar. We presented the same novel objects repeatedly to the animals for multiple days. In the session, neurons that were excited by novel objects decreased their activity as the presentation number of the repeating novel objects increased (within-day learning) while the learning rate dropped. At the start of the session on the next recording day, the majority of the activity in neurons excited by novel objects rebounded back (across-day forgetting). On average, the neurons gradually adapted to the repeating novel objects in a sawtooth pattern across days, with peaks at the start of the session and valleys at the end of the session.

It is worth noting that the neuron's across-day forgetting here is different from the recency response that we discussed earlier. The amount of forgetting reflects how much a neuron forgets about some newly learned objects overnight. On the other hand, the recency response measures the fully learned familiar objects' repetition suppression over the course of a single session. In addition, when measuring the across-day forgetting, we removed the repetition suppression effect of fully learned familiar objects. Moreover, the correlations in our data between recency

index with learning index and forgetting index were not significant. (rho = 0.028, p = 0.51, recency vs. learning, rho = 0.053, p = 0.096, recency vs. forgetting, n = 991, Spearman's correlation).

We found variety in the learning and forgetting patterns of neurons. Neurons that tend to learn more within a day also tend to forget more across days. This variety persists after grouping the neurons by brain areas. The brain areas that tend to learn more within a day also tend to forget more across days. In addition, the hippocampus and the entorhinal cortex on average have negative forgetting indices, which means that instead of forgetting, they consolidate the objects after resting at night. These two brain areas are also the key ones that participate in memory formation. (Takehara-Nishiuchi, 2014; Olafsdottir et al., 2018).

The heterogeneity of learning and forgetting further indicates that multiple systems of processing novel objects could exist. In a world where some objects are only relevant for short periods, while others must be remembered for a lifetime, heterogeneous learning systems can gate the information flow and provide adaptive behavior to the objects with various timescales. Hypothetically, the information of the objects that are only relevant for short periods would mainly stay in the fast timescale system, while only the information of the objects that are important for a long time could be gradually learned and transferred from the fast timescale system to the slow timescale system, and kept there for a long period.

4.5 Final thoughts

In this dissertation, we first studied the BF neurons' activities to salient events, including reward, uncertainty, surprise, and novelty. We found two types of neurons that process salient events in

distinct manners: one with phasic burst activity to salient events and cues predicting the events and one with ramping activity anticipating salient events. Then we studied how the brain computes novelty signals and their adaption. In multiple brain areas, we found that the computation of novelty is related to both computations of recency and sensory surprise. In addition, we found diverse timescales of neural learning and forgetting across neurons and brain areas. These results give us new insights into how the brain processes salient objects. However, there is still much work to fully understand how the brain works. We will need to combine the experimental evidence and the theoretical models to build a comprehensive framework about salience processing in the brain.

References

Adams RA, Brown HR, Friston KJ (2014) Bayesian inference, predictive coding and delusions. AVANT J Philos Int Vanguard 5:51-88.

Aitchison L, Lengyel M (2017) With or without you: predictive coding and Bayesian inference in the brain. Current opinion in neurobiology 46:219-227.

Anderson B, Mruczek REB, Kawasaki K, Sheinberg D (2008) Effects of Familiarity on Neural Activity in Monkey Inferior Temporal Lobe. Cerebral Cortex 18:2540-2552.

Apicella P, Ravel S, Deffains M, Legallet E (2011) The role of striatal tonically active neurons in reward prediction error signaling during instrumental task performance. J Neurosci 31:1507-1515.

Arendt T, Bruckner MK, Bigl V, Marcova L (1995) Dendritic Reorganization in the Basal Forebrain under Degenerative Conditions and Its Defects in Alzheimers-Disease .3. The Basal Forebrain Compared with Other Subcortical Areas. Journal of Comparative Neurology 351:223-246.

Avery MC, Krichmar JL (2017) Neuromodulatory Systems and Their Interactions: A Review of Models, Theories, and Experiments. Front Neural Circuit 11.

Avila I, Lin SC (2014) Distinct neuronal populations in the basal forebrain encode motivational salience and movement. Front Behav Neurosci 8:421.

Babayan BM, Uchida N, Gershman SJ (2018) Belief state representation in the dopamine system. Nat Commun 9:1891.

Bachman MD, Wang LL, Gamble ML, Woldorff MG (2020) Physical Salience and Value-Driven Salience Operate through Different Neural Mechanisms to Enhance Attentional Selection. Journal of Neuroscience 40:5455-5464.

Barto A, Mirolli M, Baldassarre G (2013) Novelty or surprise? Front Psychol 4:907.

Baxter MG, Chiba AA (1999) Cognitive functions of the basal forebrain. Curr Opin Neurobiol 9:178-183.

Baxter MG, Murray EA (2002) The amygdala and reward. Nature reviews neuroscience 3:563-573.

Bell AH, Summerfield C, Morin EL, Malecek NJ, Ungerleider LG (2017) Reply to Vinken and Vogels. Current Biology 27:R1212-R1213.

Berlyne DE (1950) Novelty and curiosity as determinants of exploratory behaviour. British Journal of Psychology 41:68.

Berlyne DE (1957) Uncertainty and conflict: A point of contact between information-theory and behavior-theory concepts. Psychological Review 64:329-339.

Berlyne DE (1960) Conflict, arousal, and curiosity.

Berlyne DE (1970) Novelty, Complexity, and Hedonic Value. Percept Psychophys 8:279-&.

Berns GS, Cohen JD, Mintun MA (1997) Brain regions responsive to novelty in the absence of awareness. Science 276:1272-1275.

Blackford JU, Buckholtz JW, Avery SN, Zald DH (2010) A unique role for the human amygdala in novelty detection. Neuroimage 50:1188-1193.

Bloom BH (1970) Space/Time Trade/Offs in Hash Coding with Allowable Errors. Commun Acm 13:422-&.

Bogacz R, Brown MW (2003a) Comparison of computational models of familiarity discrimination in the perirhinal cortex. Hippocampus 13:494-524.

Bogacz R, Brown MW (2003b) An anti-Hebbian model of familiarity discrimination in the perirhinal cortex. Neurocomputing 52:1-6.

Bogacz R, Brown MW, Giraud-Carrier C (2001a) Model of co-operation between recency, familiarity and novelty neurons in the perirhinal cortex. Neurocomputing 38:1121-1126.

Bogacz R, Brown MW, Giraud-Carrier C (2001b) Model of familiarity discrimination in the perirhinal cortex. J Comput Neurosci 10:5-23.

Bradley MM (2009) Natural selective attention: Orienting and emotion. Psychophysiology 46:1-11.

Bromberg-Martin ES, Hikosaka O (2011) Lateral habenula neurons signal errors in the prediction of reward information. Nat Neurosci 14:1209-1216.

Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010a) Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 68:815-834.

Bromberg-Martin ES, Matsumoto M, Nakahara H, Hikosaka O (2010b) Multiple timescales of memory in lateral habenula and dopamine neurons. Neuron 67:499-510.

Brown RE, Basheer R, McKenna JT, Strecker RE, McCarley RW (2012) Control of Sleep and Wakefulness. Physiol Rev 92:1087-1187.

Bucci DJ, Holland PC, Gallagher M (1998) Removal of cholinergic input to rat posterior parietal cortex disrupts incremental processing of conditioned stimuli. Journal of Neuroscience 18:8038-8046.

Cavanagh SE, Wallis JD, Kennerley SW, Hunt LT (2016) Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. Elife 5:e18937.

Charles DP, Gaffan D, Buckley MJ (2004) Impaired recency judgments and intact novelty judgments after fornix transection in monkeys. Journal of Neuroscience 24:2037-2044.

Cheng K, Saleem K, Tanaka K (1997) Organization of corticostriatal and corticoamygdalar projections arising from the anterior inferotemporal area TE of the macaque monkey: a Phaseolus vulgaris leucoagglutinin study. Journal of Neuroscience 17:7902-7925.

Cheung VKM, Harrison PMC, Meyer L, Pearce MT, Haynes JD, Koelsch S (2019) Uncertainty and Surprise Jointly Predict Musical Pleasure and Amygdala, Hippocampus, and Auditory Cortex Activity. Curr Biol 29:4084-4092 e4084.

Chiba AA, Bushnell PJ, Oshiro WM, Gallagher M (1999) Selective removal of cholinergic neurons in the basal forebrain alters cued target detection. Neuroreport 10:3119-3123.

Chudasama Y, Dalley JW, Nathwani F, Bouger P, Robbins TW (2004) Cholinergic modulation of visual attention and working memory: dissociable effects of basal forebrain 192-IgG-saporin lesions and intraprefrontal infusions of scopolamine. Learn Mem 11:78-86.

Costa VD, Averbeck BB (2020) Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs. Journal of Neuroscience 40:2553-2561.

Costa VD, Mitz AR, Averbeck BB (2019) Subcortical substrates of explore-exploit decisions in primates. Neuron 103:533-545. e535.

Costa VD, Dal Monte O, Lucas DR, Murray EA, Averbeck BB (2016) Amygdala and ventral striatum make distinct contributions to reinforcement learning. Neuron 92:505-517.

Cycowicz YM, Friedman D, Snodgrass JG (2001) Remembering the color of objects: an ERP investigation of source memory. Cereb Cortex 11:322-334.

Dal Monte O, Costa VD, Noble PL, Murray EA, Averbeck BB (2015) Amygdala lesions in rhesus macaques decrease attention to threat. Nature communications 6:1-10.

Dasgupta S, Sheehan TC, Stevens CF, Navlakha S (2018) A neural data structure for novelty detection. Proc Natl Acad Sci U S A 115:13093-13098.

Daye PM, Monosov IE, Hikosaka O, Leopold DA, Optican LM (2013) pyElectrode: an open-source tool using structural MRI for electrode positioning and neuron mapping. J Neurosci Methods 213:123-131.

Dotson NM, Hoffman SJ, Goodell B, Gray CM (2017) A Large-Scale Semi-Chronic Microdrive Recording System for Non-Human Primates. Neuron 96:769-782 e762.

Dotson NM, Hoffman SJ, Goodell B, Gray CM (2018) Feature-based visual short-term memory is widely distributed and hierarchically organized. Neuron 99:215-226. e214.

Dragoi V, Sharma J, Sur M (2000) Adaptation-induced plasticity of orientation tuning in adult visual cortex. Neuron 28:287-298.

Dudukovic NM, Wagner AD (2007) Goal-dependent modulation of declarative memory: neural correlates of temporal recency decisions and novelty detection. Neuropsychologia 45:2608-2620.

Egner T, Monti JM, Summerfield C (2010) Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci 30:16601-16608.

Everitt BJ, Robbins TW (1997) Central cholinergic systems and cognition. Annu Rev Psychol 48:649-684.

Fahy F, Riches I, Brown M (1993) Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. Exp Brain Res 96:457-472.

Fecteau JH, Munoz DP (2006) Salience, relevance, and firing: a priority map for target selection. Trends Cogn Sci 10:382-390.

Fell J, Klaver P, Elger CE, Fernández G (2002) The interaction of rhinal cortex and hippocampus in human declarative memory formation. Reviews in the Neurosciences 13:299-312.

Foley NC, Jangraw DC, Peck C, Gottlieb J (2014) Novelty enhances visual salience independently of reward in the parietal lobe. J Neurosci 34:7947-7957.

Friedman D, Cycowicz YM, Gaeta H (2001) The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci Biobehav R 25:355-373.

Friston K (2018) Does predictive coding have a future? Nature neuroscience 21:1019-1021.

Fukuda M, Masuda R, Ono T, Tabuchi E (1993) Responses of monkey basal forebrain neurons during visual discrimination task. Prog Brain Res 95:359-369.

Ghazizadeh A, Griggs W, Hikosaka O (2016a) Ecological Origins of Object Salience: Reward, Uncertainty, Aversiveness, and Novelty. Front Neurosci 10:378.

Ghazizadeh A, Griggs W, Hikosaka O (2016b) Object-finding skill created by repeated reward experience. Journal of Vision 16:17.

Ghazizadeh A, Fakharian MA, Amini A, Griggs W, Leopold DA, Hikosaka O (2020) Brain networks sensitive to object novelty, value, and their combination. Cerebral cortex communications 1:tgaa034.

Gottlieb J, Oudeyer P-Y, Lopes M, Baranes A (2013) Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends in cognitive sciences 17:585-593.

Gritti I, Mariotti M, Mancia M (1998) GABAergic and cholinergic basal forebrain and preoptic-anterior hypothalamic projections to the mediodorsal nucleus of the thalamus in the cat. Neuroscience 85:149-178.

Gu Z, Yakel JL (2011) Timing-dependent septal cholinergic induction of dynamic hippocampal synaptic plasticity. Neuron 71:155-165.

Hangya B, Ranade SP, Lorenc M, Kepecs A (2015) Central Cholinergic Neurons Are Rapidly Recruited by Reinforcement Feedback. Cell 162:1155-1168.

Hart EW, Jacoby J (1973) Novelty, recency, and scarcity as predictors of perceived newness. In: Proceedings of the Annual Convention of the American Psychological Association: American Psychological Association.

Hasselmo ME, Schnell E (1994) Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. J Neurosci 14:3898-3914.

Hasselmo ME, McClelland JL (1999) Neural models of memory. Current opinion in neurobiology 9:184-188.

Hasselmo ME, Sarter M (2011) Modes and models of forebrain cholinergic neuromodulation of cognition. Neuropsychopharmacology 36:52-73.

Hasselmo ME, Wyble BP, Wallenstein GV (1996) Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. Hippocampus 6:693-708.

Hasselmo ME, Fransen E, Dickson C, Alonso AA (2000) Computational modeling of entorhinal cortex. Ann N Y Acad Sci 911:418-446.

Hawco C, Lepage M (2014) Overlapping patterns of neural activity for different forms of novelty in fMRI. Front Hum Neurosci 8:699.

Hayden BY, Heilbronner SR, Pearson JM, Platt ML (2011) Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J Neurosci 31:4178-4187.

Hikosaka O, Yamamoto S, Yasuda M, Kim HF (2013) Why skill matters. Trends in cognitive sciences 17:434-441.

Homann J, Koay SA, Glidden AM, Tank DW, Berry MJ (2017) Predictive Coding of Novel versus Familiar Stimuli in the Primary Visual Cortex. biorxiv:197608.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 79:2554-2558.

Hosoya T, Baccus SA, Meister M (2005) Dynamic predictive coding by the retina. Nature 436:71-77.

Huang Y, Rao RP (2011) Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science 2:580-593.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106-154.

Jaegle A, Mehrpour V, Rust N (2019) Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. Curr Opin Neurobiol 58:167-174.

Jezzini A, Bromberg-Martin ES, Trambaiolli LR, Haber SN, Monosov IE (2021) A prefrontal network integrates preferences for advance information about uncertain rewards and punishments. Neuron 109:2339-+.

Joo HR, Frank LM (2018) The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. Nature Reviews Neuroscience 19:744-757.

Joshua M, Adler A, Bergman H (2010) Novelty encoding by the output neurons of the Basal Ganglia. Frontiers in systems neuroscience 3:20.

Kafkas A, Montaldi D (2015) The pupillary response discriminates between subjective and objective familiarity and novelty. Psychophysiology 52:1305-1316.

Kafkas A, Montaldi D (2018) How do memory systems detect and respond to novelty? Neurosci Lett 680:60-68.

Kakade S, Dayan P (2002) Dopamine: generalization and bonuses. Neural Netw 15:549-559.

Kaliukhovich DA, Vogels R (2014) Neurons in macaque inferior temporal cortex show no surprise response to deviants in visual oddball sequences. J Neurosci 34:12801-12815.

Kiehl KA, Laurens KR, Duty TL, Forster BB, Liddle PF (2001) Neural sources involved in auditory target detection and novelty processing: an event-related fMRI study. Psychophysiology 38:133-142.

Kim HF, Hikosaka O (2013) Distinct basal ganglia circuits controlling behaviors guided by flexible and stable values. Neuron 79:1001-1010.

Koay SA, Charles AS, Thiberge SY, Brody C, Tank DW (2021) Sequential and efficient neural-population coding of complex task information. bioRxiv:801654.

Kording KP, Tenenbaum JB, Shadmehr R (2007) The dynamics of memory as a consequence of optimal adaptation to a changing body. Nature neuroscience 10:779-786.

Kumaran D, Maguire EA (2007a) Match–mismatch processes underlie human hippocampal responses to associative novelty. Journal of Neuroscience 27:8517-8524.

Kumaran D, Maguire EA (2007b) Which computational mechanisms operate in the hippocampus during novelty detection? Hippocampus 17:735-748.

Kume K, Kume S, Park SK, Hirsh J, Jackson FR (2005) Dopamine is a regulator of arousal in the fruit fly. J Neurosci 25:7377-7384.

Lak A, Stauffer WR, Schultz W (2014) Dopamine prediction error responses integrate subjective value from different reward dimensions. Proc Natl Acad Sci U S A 111:2343-2348.

Laurent PA (2008) The emergence of saliency and novelty responses from Reinforcement Learning principles. Neural Networks 21:1493-1499.

Law JR, Flanery MA, Wirth S, Yanike M, Smith AC, Frank LM, Suzuki WA, Brown EN, Stark CE (2005) Functional magnetic resonance imaging activity during the gradual acquisition and expression of paired-associate memory. The Journal of neuroscience : the official journal of the Society for Neuroscience 25:5720-5729.

Ledbetter MN, Chen DC, Monosov IE (2016a) Multiple mechanisms for processing reward uncertainty in the primate basal forebrain. J Neurosci 36.

Ledbetter NM, Chen CD, Monosov IE (2016b) Multiple Mechanisms for Processing Reward Uncertainty in the Primate Basal Forebrain. J Neurosci 36:7852-7864.

Lee MG, Hassani OK, Alonso A, Jones BE (2005) Cholinergic basal forebrain neurons burst with theta during waking and paradoxical sleep. Journal of Neuroscience 25:4365-4369.

Li L, Miller EK, Desimone R (1993) The representation of stimulus familiarity in anterior inferior temporal cortex. Journal of neurophysiology 69:1918-1929.

Li YS, Nassar MR, Kable JW, Gold JI (2019) Individual neurons in the cingulate cortex encode action monitoring, not selection, during adaptive decision-making. Journal of Neuroscience 39:6668-6683.

Lin SC, Nicolelis MA (2008) Neuronal ensemble bursting in the basal forebrain encodes salience irrespective of valence. Neuron 59:138-149.

Lin SC, Brown RE, Hussain Shuler MG, Petersen CC, Kepecs A (2015) Optogenetic Dissection of the Basal Forebrain Neuromodulatory Control of Cortical Activation, Plasticity, and Cognition. J Neurosci 35:13896-13903.

Liu R, Crawford J, Callahan PM, Terry AV, Jr., Constantinidis C, Blake DT (2017) Intermittent Stimulation of the Nucleus Basalis of Meynert Improves Working Memory in Adult Monkeys. Curr Biol 27:2640-2646 e2644.

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annu Rev Neurosci 19:577-621.

Markou M, Singh S (2003) Novelty detection: a review - part 1: statistical approaches. Signal Processing 83:2481-2497.

Masuda R, Fukuda M, Ono T, Endo S (1997) Neuronal responses at the sight of objects in monkey basal forebrain subregions during operant visual tasks. Neurobiol Learn Mem 67:181-196.

Matsumoto M, Hikosaka O (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. Nature 447:1111-1115.

Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 459:837-841.

Matsumoto M, Takada M (2013) Distinct representations of cognitive and motivational signals in midbrain dopamine neurons. Neuron 79:1011-1024.

McGinty DJ, Sterman MB (1968) Sleep suppression after basal forebrain lesions in the cat. Science 160:1253-1255.

Meeter M, Murre JMJ, Talamini LM (2004) Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. Hippocampus 14:722-741.

Mehrpour V, Meyer T, Simoncelli EP, Rust NC (2021) Pinpointing the neural signatures of single-exposure visual recognition memory. Proceedings of the National Academy of Sciences 118.

Mesulam MM, Mufson EJ, Levey AI, Wainer BH (1983) Cholinergic innervation of cortex by the basal forebrain: cytochemistry and cortical connections of the septal area, diagonal band nuclei, nucleus basalis (substantia innominata), and hypothalamus in the rhesus monkey. J Comp Neurol 214:170-197.

Meyer T, Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. Proc Natl Acad Sci U S A 108:19401-19406.

Meyer T, Rust NC (2018) Single-exposure visual memory judgments are reflected in IT cortex. bioRxiv:197764.

Miljković D (2010) Review of novelty detection methods. In: The 33rd International Convention MIPRO, pp 593-598: IEEE.

Miyashita Y, Higuchi S, Sakai K, Masui N (1991) Generation of fractal patterns for probing the visual memory. Neurosci Res 12:307-311.

Monosov IE (2017) Anterior cingulate is a source of valence-specific information about value and uncertainty. Nat Commun 8:134.

Monosov IE (2020) How Outcome Uncertainty Mediates Attention, Learning, and Decision-Making. Trends in Neurosciences.

Monosov IE, Hikosaka O (2013) Selective and graded coding of reward uncertainty by neurons in the primate anterodorsal septal region. Nat Neurosci 16:756-762.

Monosov IE, Leopold DA, Hikosaka O (2015) Neurons in the Primate Medial Basal Forebrain Signal Combined Information about Reward Uncertainty, Value, and Punishment Anticipation. J Neurosci 35:7443-7459.

Monosov IE, Haber SN, Leuthardt EC, Jezzini A (2020) Anterior Cingulate Cortex and the Control of Dynamic Behavior in Primates. Current Biology 30:R1442-R1454.

Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H (2004) Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 43:133-143.

Murray EA, Wise SP (1996) Role of the hippocampus plus subjacent cortex but not amygdala in visuomotor conditional learning in rhesus monkeys. Behav Neurosci 110:1261-1270.

Murray EA, Mishkin M (1998) Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and hippocampus. Journal of Neuroscience 18:6568-6582.

Murray EA, Izquierdo A (2007) Orbitofrontal cortex and amygdala contributions to affect and action in primates. Annals of the New York Academy of Sciences 1121:273-296.

Murray EA, Baxter MG, Gaffan D (1998) Monkeys with rhinal cortex damage or neurotoxic hippocampal lesions are impaired on spatial scene learning and object reversals. Behav Neurosci 112:1291-1303.

Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D (2014) A hierarchy of intrinsic timescales across primate cortex. Nature neuroscience 17:1661-1663.

Nour MM, Dahoun T, Schwartenbeck P, Adams RA, FitzGerald THB, Coello C, Wall MB, Dolan RJ, Howes OD (2018) Dopaminergic basis for signaling belief updates, but not surprise, and the link to paranoia. Proc Natl Acad Sci U S A 115:E10167-E10176.

Nyberg L (2005) Any novelty in hippocampal formation and memory? Curr Opin Neurol 18:424-428.

Ogasawara T, Sogukpinar F, Zhang K, Feng Y-Y, Pai J, Jezzini A, Monosov IE (2022) A primate temporal cortex–zona incerta pathway for novelty seeking. Nature neuroscience 25:50-60.

Ohman A, Flykt A, Esteves F (2001) Emotion drives attention: detecting the snake in the grass. J Exp Psychol Gen 130:466-478.

Olafsdottir HF, Bush D, Barry C (2018) The Role of Hippocampal Replay in Memory and Planning. Curr Biol 28:R37-R50.

Pachitariu M, Steinmetz N, Kadir S, Carandini M, Kenneth D. H (2016) Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. biorxiv:061481.

Padoa-Schioppa C, Cai X (2011) The orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives. Ann N Y Acad Sci 1239:130-137.

Parr T, Friston KJ (2019) Attention or salience? Curr Opin Psychol 29:1-5.

Paton JJ, Buonomano DV (2018) The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. Neuron 98:687-705.

Pavlov IP (1960) Conditioned reflex: An investigation of the physiological activity of the cerebral cortex.

Pearce JM, Hall G (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87:532-552.

Peck CJ, Salzman CD (2014) The amygdala and basal forebrain as a pathway for motivationally guided attention. J Neurosci 34:13757-13767.

Peck CJ, Lau B, Salzman CD (2013) The primate amygdala combines information about space and value. Nature neuroscience 16:340-348.

Pereira JB, Hall S, Jalakas M, Grothe MJ, Strandberg O, Stomrud E, Westman E, van Westen D, Hansson O (2020) Longitudinal degeneration of the basal forebrain predicts subsequent dementia in Parkinson's disease. Neurobiol Dis 139.

Petrides M, Alivisatos B, Frey S (2002) Differential activation of the human orbital, mid-ventrolateral, and mid-dorsolateral prefrontal cortex during the processing of visual stimuli. Proc Natl Acad Sci U S A 99:5649-5654.

Pimentel MAF, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. Signal Processing 99:215-249.

Pine DS, Wise SP, Murray EA (2021) Evolution, Emotion, and Episodic Engagement. American Journal of Psychiatry:appi. ajp. 2020.20081187.

Pinto L, Goard MJ, Estandian D, Xu M, Kwan AC, Lee SH, Harrison TC, Feng G, Dan Y (2013) Fast modulation of visual perception by basal forebrain cholinergic neurons. Nat Neurosci 16:1857-1863.

Polich J (2007) Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 118:2128-2148.

Ponzi A (2008) Dynamical model of salience gated working memory, action selection and reinforcement based on basal ganglia and dopamine feedback. Neural Netw 21:322-330.

Preuschoff K, t Hart BM, Einhauser W (2011) Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. Front Neurosci 5:115.

Puglisi-Allegra S, Ventura R (2012) Prefrontal/accumbal catecholamine system processes high motivational salience. Front Behav Neurosci 6:31.

Radulescu A, Niv Y, Ballard I (2019) Holistic Reinforcement Learning: The Role of Structure and Attention. Trends Cogn Sci 23:278-292.

Ramachandran S, Meyer T, Olson CR (2016) Prediction suppression in monkey inferotemporal cortex depends on the conditional probability between images. Journal of neurophysiology 115:355-362.

Ranganath C, Rainer G (2003) Neural mechanisms for detecting and remembering novel events. Nat Rev Neurosci 4:193-202.

Raver SM, Lin SC (2015) Basal forebrain motivational salience signal enhances cortical processing and decision speed. Front Behav Neurosci 9:277.

Redgrave P, Gurney K (2006) The short-latency dopamine signal: a role in discovering novel actions? Nature Reviews Neuroscience 7:967-975.

Reichardt R, Polner B, Simor P (2020) Novelty manipulations, memory performance, and predictive coding: The role of unexpectedness. Frontiers in human neuroscience 14:152.

Rescorla RA (2004) Spontaneous recovery. Learning & Memory 11:501-509.

Richardson RT, DeLong MR (1990) Context-dependent responses of primate nucleus basalis neurons in a go/no-go task. J Neurosci 10:2528-2540.

Roesch MR, Calu DJ, Esber GR, Schoenbaum G (2010) All that glitters ... dissociating attention and outcome expectancy from prediction errors signals. J Neurophysiol 104:587-595.

Russchen FT, Amaral DG, Price JL (1985) The afferent connections of the substantia innominata in the monkey, Macaca fascicularis. J Comp Neurol 242:1-27.

Sarafyazd M, Jazayeri M (2019) Hierarchical reasoning by neural circuits in the frontal cortex. Science 364:eaav8911.

Saunders RC, Murray EA, Mishkin M (1984) Further evidence that amygdala and hippocampus contribute equally to recognition memory. Neuropsychologia 22:785-796.

Saunders RC, Mishkin M, Aggleton JP (2005) Projections from the entorhinal cortex, perirhinal cortex, presubiculum, and parasubiculum to the medial thalamus in macaque monkeys: identifying different pathways using disconnection techniques. Exp Brain Res 167:1-16.

Schomaker J, Meeter M (2015) Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. Neurosci Biobehav R 55:268-279.

Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophysiol 80:1-27.

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241-263.

Schultz W (2016) Dopamine reward prediction error coding. Dialogues Clin Neurosci 18:23-32.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593-1599.

Schwartenbeck P, FitzGerald T, Dolan R, Friston K (2013) Exploration, novelty, surprise, and free energy minimization. Frontiers in psychology 4:710.

Shuler MG, Bear MF (2006) Reward timing in the primary visual cortex. Science 311:1606-1609.

Smith MA, Ghazizadeh A, Shadmehr R (2006) Interacting adaptive processes with different timescales underlie short-term motor learning. PLoS biology 4:e179.

Soltani A, Izquierdo A (2019) Adaptive learning under expected and unexpected uncertainty. Nature Reviews Neuroscience 20:635-644.

Spitmaan M, Seo H, Lee D, Soltani A (2020) Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. Proceedings of the National Academy of Sciences 117:22522-22531.

Stefanacci L, Suzuki WA, Amaral DG (1996) Organization of connections between the amygdaloid complex and the perirhinal and parahippocampal cortices in macaque monkeys. The Journal of comparative neurology 375:552-582.

Stickgold R (2005) Sleep-dependent memory consolidation. Nature 437:1272-1278.

Strange BA, Dolan RJ (2001) Adaptive anterior hippocampal responses to oddball stimuli. Hippocampus 11:690-698.

Strange BA, Duggins A, Penny W, Dolan RJ, Friston KJ (2005) Information theory, novelty and hippocampal responses: unpredicted or unpredictable? Neural Networks 18:225-230.

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction: MIT press.

Sutton S, Braren M, Zubin J, John ER (1965) Evoked-potential correlates of stimulus uncertainty. Science 150:1187-1188.

Suzuki WA (1996) The anatomy, physiology and functions of the perirhinal cortex. Current opinion in neurobiology 6:179-186.

Szymusiak R, McGinty D (1986) Sleep suppression following kainic acid-induced lesions of the basal forebrain. Exp Neurol 94:598-614.

Takahashi YK, Langdon AJ, Niv Y, Schoenbaum G (2016) Temporal Specificity of Reward Prediction Errors Signaled by Putative Dopamine Neurons in Rat VTA Depends on Ventral Striatum. Neuron 91:182-193.

Takahashi YK, Batchelor HM, Liu B, Khanna A, Morales M, Schoenbaum G (2017) Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. Neuron 95:1395-1405 e1393.

Takehara-Nishiuchi K (2014) Entorhinal cortex and consolidated memory. Neurosci Res 84:27-33.

Tapper AR, Molas S (2020) Midbrain circuits of novelty processing. Neurobiology of Learning and Memory:107323.

Tiitinen H, May P, Reinikainen K, Näätänen R (1994) Attentive novelty detection in humans is governed by pre-attentive sensory memory. Nature 372:90.

Tulving E, Kroll N (1995) Novelty assessment in the brain and long-term memory encoding. Psychon Bull Rev 2:387-390.

Tulving E, Markowitsch HJ, Craik FE, Habib R, Houle S (1996) Novelty and familiarity activations in PET studies of memory encoding and retrieval. Cereb Cortex 6:71-79.

Turchi J, Chang C, Frank QY, Russ BE, David KY, Cortes CR, Monosov IE, Duyn JH, Leopold DA (2018) The basal forebrain regulates global resting-state fMRI fluctuations. Neuron 97:940-952. e944.

Tyulmankov D, Yang GR, Abbott L (2021) Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. Neuron.

Tyulmankov D, Yang GR, Abbott LF (2022) Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. Neuron 110:544-557 e548.

Utzerath C, Schmits IC, Buitelaar J, de Lange FP (2018) Adolescents with autism show typical fMRI repetition suppression, but atypical surprise response. Cortex 109:25-34.

Võ MLH, Jacobs AM, Kuchinke L, Hofmann M, Conrad M, Schacht A, Hutzler F (2008) The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. Psychophysiology 45:130-140.

Vogels R (2016) Sources of adaptation of inferior temporal cortical responses. Cortex 80:185-195.

Voytko ML (1996) Cognitive functions of the basal forebrain cholinergic system in monkeys: memory or attention? Behav Brain Res 75:13-25.

Voytko ML, Olton DS, Richardson RT, Gorman LK, Tobin JR, Price DL (1994) Basal forebrain lesions in monkeys disrupt attention but not learning and memory. J Neurosci 14:167-186.

Waite JJ, Wardlow ML, Power AE (1999) Deficit in selective and divided attention associated with cholinergic basal forebrain immunotoxic lesion produced by 192-saporin; Motoric/sensory deficit associated with Purkinje cell immunotoxic lesion produced by OX7-saporin. Neurobiology of Learning and Memory 71:325-352.

Wallis JD, Rich EL (2011) Challenges of Interpreting Frontal Neurons during Value-Based Decision-Making. Front Neurosci 5:124.

Wang T, Mitchell CJ (2011) Attention and relative novelty in human perceptual learning. J Exp Psychol Anim Behav Process 37:436-445.

Wessel JR, Danielmeier C, Morton JB, Ullsperger M (2012) Surprise and error: common neuronal architecture for the processing of errors and novelty. Journal of Neuroscience 32:7528-7537.

White JK, Monosov IE (2016) Neurons in the primate dorsal striatum signal the uncertainty of object-reward associations. Nat Commun 7:12735.

White JK, Bromberg-Martin ES, Heilbronner SR, Zhang K, Pai J, Haber SN, Monosov IE (2019) A neural network for information seeking. Nat Commun 10:5168.

Whitehouse PJ, Price DL, Struble RG, Clark AW, Coyle JT, Delong MR (1982) Alzheimers-Disease and Senile Dementia - Loss of Neurons in the Basal Forebrain. Science 215:1237-1239.

Wildenberg GA, Rosen MR, Lundell J, Paukner D, Freedman DJ, Kasthuri N (2021) Primate neuronal connections are sparse in cortex as compared to mouse. Cell Rep 36:109709.

Wilson FA, Rolls ET (1990) Neuronal responses related to reinforcement in the primate basal forebrain. Brain Res 509:213-231.

Wilson FA, Ma YY (2004) Reinforcement-related neurons in the primate basal forebrain respond to the learned significance of task events rather than to the hedonic attributes of reward. Brain Res Cogn Brain Res 19:74-81.

Xiang J-Z, Brown M (1998) Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. Neuropharmacology 37:657-676.

Xu M, Chung S, Zhang S, Zhong P, Ma C, Chang WC, Weissbourd B, Sakai N, Luo L, Nishino S, Dan Y (2015) Basal forebrain circuit for sleep-wake control. Nat Neurosci 18:1641-1647.

Yamaguchi S, Hale LA, D'Esposito M, Knight RT (2004) Rapid prefrontal-hippocampal habituation to novel events. Journal of Neuroscience 24:5356-5363.

Yamamoto S, Monosov IE, Yasuda M, Hikosaka O (2012) What and where information in the caudate tail guides saccades to visual objects. The Journal of neuroscience : the official journal of the Society for Neuroscience 32:11005-11016.

Yasuda M, Yamamoto S, Hikosaka O (2012) Robust representation of stable object values in the oculomotor Basal Ganglia. J Neurosci 32:16917-16932.

Yu B, Lu J, Li X, Zhou J (2021) Salience-Aware Face Presentation Attack Detection via Deep Reinforcement Learning. IEEE Transactions on Information Forensics Security.

Zaborszky L, Csordas A, Mosca K, Kim J, Gielow MR, Vadasz C, Nadasdy Z (2015) Neurons in the basal forebrain project to the cortex in a complex topographic organization that reflects corticocortical connectivity patterns: an experimental study based on retrograde tracing and 3D reconstruction. Cereb Cortex 25:118-137.

Zaborszky L, Gombkoto P, Varsanyi P, Gielow MR, Poe G, Role LW, Ananth M, Rajebhosale P, Talmage DA, Hasselmo ME, Dannenberg H, Minces VH, Chiba AA (2018) Specific Basal Forebrain-Cortical Cholinergic Circuits Coordinate Cognitive Operations. J Neurosci 38:9446-9458.

Zhang K, Chen CD, Monosov IE (2019) Novelty, Salience, and Surprise Timing Are Signaled by Neurons in the Basal Forebrain. Curr Biol 29:134-142 e133.

Zhu Y, Nachtrab G, Keyes PC, Allen WE, Luo L, Chen X (2018) Dynamic salience processing in paraventricular thalamus gates associative learning. Science 362:423-429.