UNIVERSITY OF TRENTO

Doctoral Programme in Cognitive Science

THE ROLE OF TASK RELEVANCE IN THE MODULATION OF BRAIN DYNAMICS DURING SENSORY PREDICTIONS

ANTONINO GRECO

XXXIII PhD Cycle

Advisor: Andrea Caria

May 2021

Abstract

Associative learning is a fundamental ability biological systems possess in order to adapt to a nonstationary environment. One of the core aspects of associative learning theoretical frameworks is that surprising events drive learning by signalling the need to update the system's beliefs about the probability structure governing stimuli associations. Specifically, the central neural system generates internal predictions to anticipate the causes of its perceptual experience and compute a prediction error to update its generative model of the environment, an idea generally known as the predictive coding framework. However, it is not clear whether the brain generates these predictions only for goal-oriented behavior or they are more a general characteristic of the brain function. In this thesis, I explored the role of task relevance in modulating brain activity when exposed to sensory associative learning task. In the first study, participants were asked to perform a perceptual detection task while audio-visual stimuli were presented as distractors. These distractors possessed a probability structure that made some of them more paired than others. Results showed that occipital activity triggered by the conditioned stimulus was elicited just before the arrival of the unconditioned visual stimulus. Moreover, occipital activity after the onset of the unconditioned stimulus followed a pattern of precision-weighted prediction errors. In the second study, two more sessions were added to the task in the previous study in which the probability structure for all stimuli associations was identical and the whole experiment was spanned in six days across two weeks. Results showed a difference in the modulation of the beta band induced by the presentation of the unconditioned stimulus preceded by the predictive and unpredictive conditioned auditory stimuli by comparing the pre and post sessions activity. In the third study, participant were exposed to a similar task respect to the second study with the modification that there was a condition in which the conditioned-uncoditioned stimulus association was task-

relevant, thus allowing to directly compare task-relevant and task-irrelevant associations. Results showed that both types of associations had similar patterns in terms of activity and functional connectivity when comparing the brain responses to the onset of the unconditioned visual stimulus. Taken togheter, these findings demonstrate irrelevant associations rely on the same neural mechanisms of relevant ones. Thus, even if task relevance play a modulatory role on the strenght of the neural effects of associative learning, predictive processes take place in sensory associative learning regardless of task relevance.

Acknowledgements

Throughout my years at the Department of Psychology and Cognitive Science, my supervisor Andrea Caria gave me an extraordinary combination of support, guidance and advice for which I am immensely grateful. His comments and suggestions never failed to be insightful and shaped my views on science and the sort of research I intend to pursue.

I am also thankful to all my friends and colleagues who helped me to build the skills and knowledge I needed for the completion of this work. A special thank to Marco and Giovanni for the discussions about the computational models and their mathematical details and to Christoph for the fruitful discussions about signal processing methods for neuroimaging data.

A very special thank to my beloved wife Clara. Without you, this and much more would never be possible.

Finally, I do not know how to express my gratitude to my family for their support and love. Grazie di tutto, mamma e papà.

Table of contents

List of figures

Thesis Outline and Chapter Summary

The aim of this thesis was to assess the role of task relevance in the modulation of brain dynamics during sensory associative learning, using a combination of computational models of associative learning and multivariate pattern analysis with machine learning models applied to the collected neuroimaging data (Electroencephalography and Magnetoencephalography). A range of associative learning tasks was used with different protocols of classical conditioning procedures and probability structures. This thesis is organized as follows:

Chapter 1: General Introduction – In the first chapter, there is a brief overview of the field of associative learning, starting from the early works of Pavlov and discusses crucial experiments that influenced the current theoretical view of the mechanisms underlying associative learning. In particular, the focus is on the role of the stimulus-stimulus contingency on the associability between two stimuli or, in other words, how prediction and surprise drive associative learning. In addition, there is a review of the major evidence about the neural mechanisms of associative learning from the micro level, such as the role of the dopamine neurotrasmitter, to the macro level, such as how brain areas interact during sensory associations. Finally it is introduced some literature about how task relevance influences stimulus-stimulus associations.

Chapter 2: Stimulus-independent visual cortex activity induced by implicit auditory conditioning – In this chapter, there is reported the first study of this thesis about the modulation of the visual cortex by auditory activity during task-irrelevant associative learning. Participants were asked to perform a detection task while audio-visual stimuli were presented as distractors. These distractors possessed a probability structure that made some of them more paired than others. Results showed that participants learned these task-irrelevant associations even without being aware of them. Moreover, we observed an occipital activity triggered by the conditioned auditory stimulus just before the arrival of the visual outcome and that occipital activity after the onset of the unconditioned visual stimulus followed a pattern of precision-weighted prediction errors estimated using an ideal Bayesian observer computational model.

Chapter 3: Task-irrelevant sensory associations modulate visual oscillatory activity in the beta band – In this chapter, there is reported the second study in which we investigated time-frequency representations of the EEG signal underlying task-irrelevant associations. We presented to the participants audio-visual associations while performing a perceptual detection task, thus intentionally directing their attention away from the audio-visual associations and making them irrelevant for the task they were instructed to perform (same as the first study). In this study, we added two more sessions in which the probability structure for all stimuli associations was identical before and after the main task and spanned the whole experiment in six days across two weeks. We found that participants learned these associations without being aware, confirming the findings of the first study. The key finding of this study was a difference in the modulation of the beta band induced by the presentation of the

unconditioned visual stimulus preceded by the predictive and unpredictive conditioned auditory stimuli by comparing the pre and post sessions activity. Therefore, we demonstrated that task-irrelevant associations are captured by the brain even when spread across a long time range such as in this experiment.

Chapter 4: Revealing the similarity of relevant and irrelevant associations induced brain dynamics – In this chapter, there is reported the third study in which we investigated the time-locked activity and functional connectivity networks of the MEG signal underlying task-relevant and task-irrelevant associations. Participants were exposed to an audio-visual stream of stimuli while performing a perceptual detection task in which they had to press a button when perceiving the visual target, thus intentionally directing their attention to the cue-target association and making the other audio-visual associations irrelevant for the task they were instructed to perform. One of these pairings had the same probability structure of the cue-target association across the experimental sessions, while the other had a uniform probability structure across the entire experiment. Results showed that relevant and irrelevant associations had similar patterns of activation when comparing the brain responses to the onset of the visual stimulus. This can be interpreted as evidence that prediction errors are computed similarly regardless of the task relevance.

Chapter 5: General Discussion – This chapter provides a general discussion and the conclusions of this work, presenting its contributions to the field of associative learning as well as it limitations and suggests directions for future research.

A brief history of Associative Learning

Learning is the ability to acquire, store and retrieve information. It is nowadays considered the hallmark of cognition since every physical system, either biological or artificial, that is able to learn is regarded as a cognitive system. In the scientific community studying how to provide cognitive abilities to artificial systems, the importance of learning was only recently recognized. Now the majority of the researchers in that field consider learning as the fundamental step to achieve artificial general intelligence (Jordan & Mitchell, 2015). On the other hand, the study of learning in biological systems has more than a century of scientific research. It has been one of the central fields in disciplines such as Psychology and Cognitive Neuroscience since their foundation. In biological systems, the process of learning is almost ubiquitously undertaken by the nervous system. In particular, in recently evolved nervous systems there are some neurons specialized to handle specific aspects of the learning process. The peripheral nervous system, together with the sensory areas of the central nervous system, is mainly assigned to the process of acquiring information from the external world and also from the internal states of the organism, a process referred to as perception. Contrarily to the process of perception that is largely known and well understood, the ability to store the acquired information is not so clear. The traditional view of the neuroscientific community on this topic is that the information is encoded in the strength of the synapses connecting all the neurons in the brain. This theory has recently been debated by some empirical evidence that posit the ramification of the dendrites as an encoding mechanism that can be complementary to the synaptic mechanism (Leuner et al., 2003; Ryan et al., 2015). Nevertheless, a robust finding of the literature on memory abilities is that the hippocampus, a brain structure embedded deep in the

temporal lobe of each cerebral cortex, is crucially responsible for the encoding of the information, especially for the long-term stability of the encoding. One of the key evidence in favor of this notion is the case of Henry Gustav Molaison, also known as H.M. (Squire, 2009), who had a bilateral medial temporal lobectomy to surgically remove the anterior two-thirds of his hippocampi, parahippocampal cortices, entorhinal cortices, piriform cortices, and amygdalae to cure an intractable form of epilepsy. After the surgery, H.M. developed a severe anterograde amnesia. He was completely unable to form new semantic knowledge. He had also mild retrograde amnesia meaning that he could not remember most of the memories encoded up to two years before the surgery. Finally, the ability to retrieve the encoded information is mainly understood as the process of functionally reactivating the neural pattern storing that information. This is also generally orchestrated by the hippocampus, although the structure of the connectome plays a significant role in the efficiency of the retrieval procedure (Brodt et al., 2018; Frankland et al., 2019). Therefore, if two neural patterns representing two different information are encoded in large-scale brain networks that are very interconnected between each other, the act of retrieving one information will be more efficient if the other one is already functionally activated. Historically, the study of learning in animals and humans has been divided into non-associative and associative learning. The former refers to a change in the strength of response to a stimulus due to repeated exposure. Non-associative learning can be distinguished in habituation and sensitization, two terms usually used to denote, respectively, the decrease and the increase of the response. Associative learning refers to the process of establishing an association between two or more events or stimuli (Delamater & Matthew Lattal, 2014; Pearce & Bouton, 2001). It encompasses, in practice, most of the learning phenomena that we encounter in everyday life or experimental settings. One of the first researchers that started to investigate

systematically associative learning was Edward Thorndike. He, in his 1898 dissertation on animal intelligence, proposed a theory of associative learning in animals, the so-called 'law of effect' (Thorndike, 1898), advocating that learning involves the establishment of associations that are constituted when responses are followed by rewards. In the same period, Ivan Pavlov, a Russian physiologist, was conducting an experiment in dogs to study the digestive system and the chemical composition of saliva. He accidentally discovered that after some time he delivered the food to the dog, the dog started to salivate before the presentation of the food (Pavlov, 1927). Upon closer examination, he realized that actually, the dog was salivating when his assistant was ringing a bell to indicate that the food was ready. Surprised by this phenomenon, he left the investigation of the digestive system and started to study this "psychic secretion" response, as he termed it. This new line of research was termed classical conditioning or Pavlovian conditioning, in his honour. In Pavlovian conditioning, a biologically salient unconditioned stimulus (US, often also termed "reinforcer") such as the food delivery, elicits an unconditioned response (UR), the salivation (Pavlov, 1927). When a neutral conditioned stimulus (CS) such as the ring of the bell regularly precedes the US, the CS will eventually also elicit salivation as a conditioned response (CR). This form of associative learning was conceived as a stimulus-stimulus association, to distinguish it from operant or instrumental conditioning in which the association is between a stimulus and an action selected by the organism, i.e. a stimulus-response association. The importance of stimulus-stimulus associations was immediately recognized by the scientific community that studied animal and human behavior at that time and the study of classical conditioning immensely flourished. One of the fundamental research questions that was addressed in the first experiments was to determine the necessary and sufficient conditions under

which two stimuli are associated. At first, it seemed reasonable to postulate that the temporal contiguity of the CS and the US was a necessary and sufficient

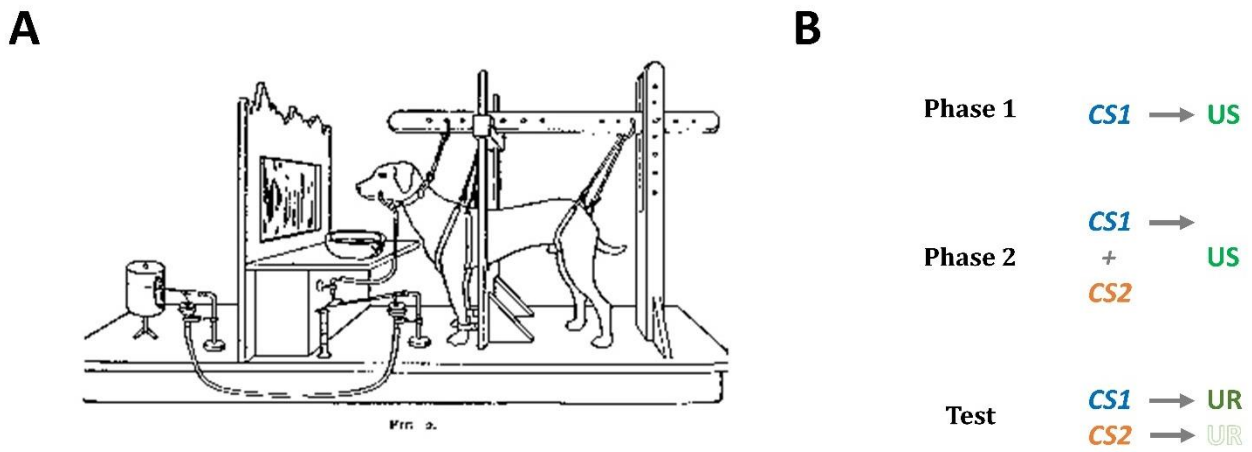

FIGURE 1 A. Figure adapted from Yerkes and Morgulis (1909), showing the setup for the Pavlov's experiments on the conditional reflexes. **B.** Abstract scheme of the blocking paradigm showing that when a CS is paired with an US simultaneously with a precedingly paired CS, the latter CS is not able to elicit a conditioned response.

condition for associative learning taking place. But in 1969, Kamin (1969) demonstrated that this was not the case, showing a characteristic phenomenon of classical conditioning known as "blocking". In the first session of a blocking paradigm, an initially neutral CS (A) is paired with an US, and another neutral CS (B) is presented but never paired. After this session, A will evoke a CR, but B will not. In a second session, A is presented in combination with another CS (X), and B with another CS (Y), and both compound CSs are repeatedly associated with the US. After the second session, Y will evoke a CR, whereas X will not, even though both CSs have been paired with a US equally often. This can be explained by noting that for the AX compound, the US could be

fully predicted by A alone, rendering X redundant, whereas for the BY compound, B could not anticipate the US, leaving it available to be associated to Y. This suggests that when an US is completely predicted by the CSs, no further learning occurs. In other words, A had "blocked" the formation of an association between X and the US. In the same period, Robert Rescorla demonstrated through a series of experiments that the contingency, the predictability of the US given the CS, was an essential requirement for Pavlovian conditioning. In other words, the CS and the US become associated if and only if the CS carries information about the presence or the absence of the US. In a critical experiment, Rescorla (1968) trained a sample of rats to press a lever in an experimental chamber to get a food pellet for 2 hours. Then there were five sessions in which the lever was blocked. In each of these sessions, 12 tones with a duration of 2 minutes were presented at random intervals. During these sessions, the rats were also occasionally exposed to very short, mildly painful electric shocks to their feet. Rescorla manipulated the distribution of the shocks relative to the tones in three different groups. For one group, 12 shocks per session were absolutely contingent on the tone, such as they only occurred when it was on. The rats in a second control group, also received 12 shocks for each session while the tone was on, but they also received shocks at a frequency of 30 seconds during the time when the tone was not presented. In this second group, the number or frequency of tone-shock associations was not altered, but the tone-shock contingency was strongly decreased with respect to the first group and also the total number of shocks per session was greatly increased (Rescorla, 1968). To control also for these changes, Rescorla ran a third control group in which the subjects received 12 shocks, the same total as the first group, but distributed at a truly random rate without any regard to the tone. In order to test the magnitude of the association between the tone and the shock in the different groups, Rescorla first removed the rats' fear of the experimental

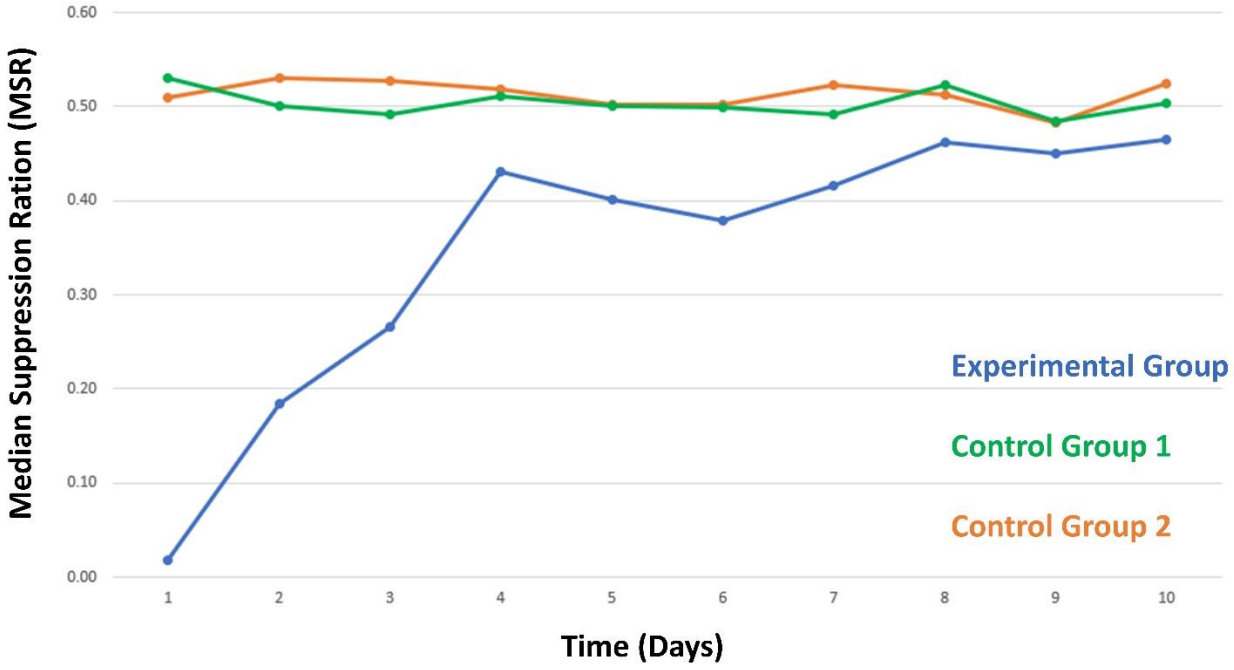

FIGURE 2. Data adapted from Rescorla (1968). Results are plotted in terms of a suppression ratio of the form A/(A + B) where A is the rate of responding in CS and B is the rate of responding in a comparable period prior to CS onset. Thus, a suppression ratio of 0 indicates no responding during CS while one of 0.5 indicates similar rates of responding during CS and the pre-CS period.

chamber, with additional two sessions in which the lever was available to use and there were not any tones nor shocks. After these additional sessions, the subjects restarted pressing the lever for food. In the last sessions, the experimenter measured the conditioned fear of the rats by tracking their willingness to press the lever when the tone was presented. If the presentation of the tone made the rats afraid of it, they froze until the tone is presented and then they resumed to press the lever. The results were that the rats in the first group learned to fear the tone, but the others in both the two control groups did

not (Fig. 2). Thus, Rescorla concluded that it is the CS-US contingency and not temporal contiguity that drives Pavlovian conditioning (Rescorla, 1968). To further demonstrate the central role of contiguity, another example that can be considered is from the protocol of inhibitory conditioning. This experimental procedure is termed inhibitory to differentiate from canonical excitatory conditioning, in which the US is followed by the CS. In the simplest inhibitory protocol, the US occurs only when the CS is absent (Rescorla, 1966). Although it is a popular experimental paradigm among the researcher in the field, it has been often underestimated the implication from the fact that in this procedure an association is formed between the CS and the absence of the US. In other words, the organism learns an association by systematically not pairing the CS with the US, thus precluding every consideration about the temporal contiguity because in this case there cannot be contiguity between an event and a "non-event" (Gallistel, 2002). To summarize the discussed literature, associative learning takes place only when the CS is informative about some characteristics of the US such as the timing of arrival or its absence, or in other words, the presence of the CS reduces the uncertainty of the organism about some features of the US. Therefore, associative learning is almost entirely driven by how surprising a stimulus-stimulus association is. Specifically, the more a CS-US association is surprising for the biological system, the more will be strengthened. In recent years, this notion of surprise-driven associative learning became more and more relevant and nowadays encompasses the majority of associative learning models (Smith et al., 2006; Terao et al., 2015). The crucial mechanism that is postulated is that an organism constantly compares the predictions made by its internal model of the world with the gathered sensory data and updates the model accordingly, based on the mismatch between the predicted and actual outcome (Clark, 2013). The surprise is formally defined as the difference between the predicted and actual outcome and this delta is often

referred as prediction error (Rao & Ballard, 1999). In the next section, it will follow a general discussion about the neural mechanisms underlying Pavlovian conditioning and the accumulating neurobiological evidence about how the brain computes these prediction errors.

Neural basis of Pavlovian Conditioning

Pavlovian conditioning is a basic form of associative learning usually regarding the association of two perceptual stimuli. In 1949, Donald Hebb suggested that this association is encoded in the strength of the synapses connecting the neurons that store the two stimuli's representation (Hebb, 1949), a concept that was summarized with the term synaptic plasticity. Following this conjecture, researchers actually found some molecular mechanisms that corroborate the Hebb's idea. What is known is that synaptic plasticity is mediated by N-methyl-D-aspartate (NMDA) receptors, which modulate the number of α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor expressed at the synapse (Genoux & Montgomery, 2007). Presynaptic neuromodulator release in co-occurrence with postsynaptic depolarisation enables a calcium influx through the NMDA receptors, which induces trafficking and the phosphorylation of glutamatergic AMPA receptors. These characteristics of NMDA receptors make them ideally eligible for associative learning processes that entail concurrent activity in different areas of the nervous system through the general process of spike-timing-dependent-plasticity (STDP, Markram et al., 1997). Although NMDA-dependent mechanisms have been found to play a key role in learning and memory processes in the brain (Ji et al., 2005; Tye et al., 2008), there are some recent studies that posit the morphological structure of the dendritic arborisation or dendritic spines as an essential element that regulates associative learning mechanisms in conjunction with synaptic

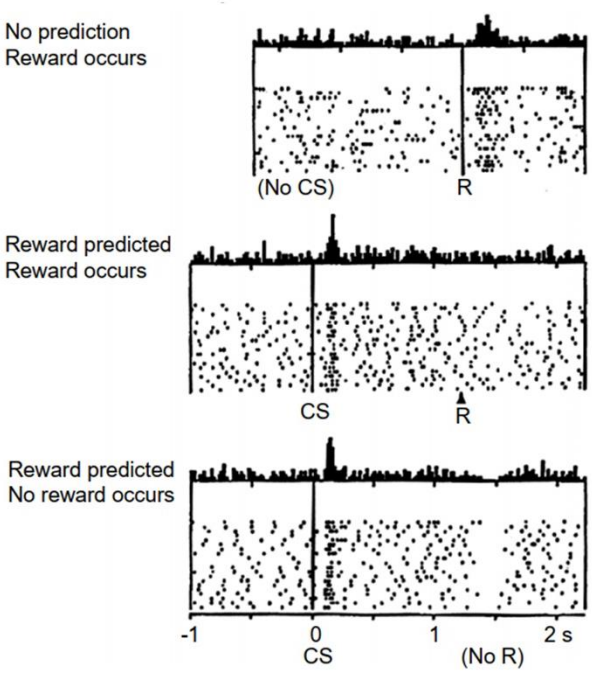

FIGURE 3. Figure adapted from Schultz et al. (1997). Changes in dopamine neuron firing reflect the prediction errors of appetitive events. For each panel, the top graph represents the accumulated spike count per time bin, and each dotted line in represents one recording session, where each dot is a spike. Before learning, the juice drop is not predicted, resulting in a positive prediction error, and increased firing in response to the reward. After learning, the CS predicts the reward, and the dopamine neurons increase firing rate in response to the predictive CS, but not to the predicted reward. When after learning the CS is presented, but the reward is omitted, this results in a negative prediction error and suppressed firing of the DA neurons at the time the reward should have occurred.

plasticity (Tazerart et al., 2020). For example, Bencsik et al. (2019) showed that calcium/calmodulin-dependent serine protein kinase (CASK) interactive proteins, multidomain neuronal scaffold proteins such as Caskin1 and Caskin2, influenced the learning capabilities of mice via regulating dendritic spine morphology and AMPA receptor localisation. Another key factor that contributes to changes in the synaptic strength and the morphology of dendritic arborisation is the dopamine (DA) neurotransmitter. There is an extensive

literature that shows how DA regulate the trafficking, insertion, phosphorylation and endocytosis of NMDA receptors (Jiao et al., 2007; Salazar-Colocho et al., 2007) as well as the formation of dendritic spines (Fasano et al., 2013). It is also well known the dopaminergic system is responsible for encoding the prediction errors or the surprise for the rewarding stimuli in a classical conditioning setting. It has been shown that when salient stimuli are presented to monkeys, their DA neurons in the ventral striatum firmly increase their firing rate (Tanaka et al., 2019). These salient stimuli can be biologically relevant assets such as food and water, but also any other stimuli carrying information about the arrival of these goods (Ljungberg et al., 1992; Romo & Schultz, 1990). Waelti et al. (2001) also showed that these DA responses were conformed to the behavioral pattern found in the blocking paradigms. In an influential series of studies, Schultz and colleagues investigated the phasic DA firing pattern during Pavlovian conditioning in the macaque ventral tegmental area (VTA, Mirenowicz & Schultz, 1994, 1996; Romo & Schultz, 1990; Schultz, 1998). When a neutral visual stimulus is presented to a primate followed by a juice reward, the DA neurons increase firing in response to the reward, but not in response to the visual stimulus. As the primate learns the CS-US association, firing rates increase when the CS is presented. Once the association is learned, the US triggers progressively smaller increases in firing. When the US is completely predicted, firing rates stop increasing, while when the US is omitted, firing rates decrease to below baseline. This pattern of firing rates indicates that what the DA neurons react to is not the US itself, but its prediction error. Subsequently, fMRI studies on humans found VTA also encodes reward prediction errors (Bray & O'Doherty, 2007; Colas et al., 2017; D'Ardenne et al., 2008; Kumar et al., 2018). The involvement of the VTA can be explained by the fact that the ventral striatum receives dopaminergic projections from the VTA (Joel & Weiner, 2000) and the BOLD signal reflects

more postsynaptic potentials than firing rate, thus it can tell more about the input on an area rather than an output (Logothetis et al., 2001). Recent findings also showed amygdala and frontal cortex responsible for reward prediction error encoding, but only in subpopulations of neurons (Schultz, 2016). Concerning the functional aspects of the neural mechanisms involved in Pavlovian conditioning, the way in which dopamine activity encodes CS-US associations has been extensively studied both theoretically and experimentally. The model-free reinforcement learning algorithm described by Sutton and Barto (1981) has been successful in modelling the phasic activity of the dopaminergic midbrain system as well as in other cortical regions (O'Doherty et al., 2003). In this algorithm, the discrepancies between the expected and delivered outcome are computed over consecutive time steps during a trial. Crucially, the prediction error signal usually elicited by the reward is transferred temporally back to the stimulus that reliably predicts reward delivery. This effectively assigns to a cue that predicts the reward the value inherent in the reward itself, rather than just encoding the occurrence of the reward (Sutton & Barto, 1981). In recent times, all the experimental evidence described above led to the creation of a general theoretical framework capable of accounting nearly any observed phenomenon about associative learning and brain function in general. It is called Predictive Coding (PC), and it encompasses a family of theoretical constructs about how the brain works, such as the Bayesian brain hypothesis and the free-energy principle (Clark, 2013; Friston et al., 2006; Friston, 2010; Hawkins & Blakeslee, 2004; Huang & Rao, 2011; Mumford, 1992; Rao & Ballard, 1999). According to the PC framework, the brain is essentially a hierarchical prediction machine (Clark, 2013). The brain is constantly confronted with a great abundance of sensory information that must be processed efficiently to produce appropriate behavioural outcomes. One way of optimizing this process is to predict incoming sensory information based on experience so that expected information

is processed efficiently and computational resources can be allocated accordingly. PC argues that the nervous system constantly generates models of the world based on contextual and stored information (Friston, 2010). Such predictive model is implemented in higher cortical areas and transmitted through feedback connections to lower sensory cortices (Friston et al., 2006). Conversely, feedforward connections process the mismatch between the predicted information and the actual sensory input (Rao & Ballard, 1999). The predictive model is constantly updated according to this prediction error signal. The origins of this idea go back to work on the perception of von Helmholtz, who was the first to conceptualize perception as a process of probabilistic, knowledge-driven inference (Helmholtz, 1925). Helmholtz's key idea was that "sensory systems are in the tricky business of inferring sensory causes from their bodily effects" (Clark, 2013). In order to accomplish that, it is required computing multiple probability distributions, because a single such effect will be coherent with various sets of causes differentiated just by their relative probability of occurrence. One of the most established models is surely the free-energy principle (Friston et al., 2006). The free-energy principle is based on considerations about the thermodynamics of living organisms. The main problem to be addressed for biological systems is to maintain stable their structure in spite of the continuous change of the environment, due to the fact that the repertoire of physiological states in which an organism can survive is limited. This implies that the probability of these sensory states must have a low entropy associated, and the notion of entropy in information theory is equivalent to the concept of surprise. The more entropy is high, the more sensory data will be unexpected, and for a biological system that means being in danger. Therefore, a biological system has to "minimize the long-term average of surprise to ensure that their sensory entropy remains low" (Friston, 2010). In a thermodynamic sense, free energy is a measure of the energy available to do

useful work. When this concept is applied to the study of cognition, free-energy emerges as the difference between the way the world is represented in neural circuits, and the way it actually is. This means for a biological system that minimizing free energy implies the reduction of surprise. Thus, a biological agent may suppress free energy only by changing the two things it depends on: they can change sensory input by acting on the world or they can change their recognition density by changing their internal states. These two processes are isomorphic to the concepts of perception and action, and this implies that free energy principle prescribes the optimal conditions for the realization of the perception-action cycle. One of the most basic and robust paradigms to demonstrate predictive processing activity in the brain stimuli is the oddball paradigm (Näätänen et al., 1978). In this experimental procedure, the presentation of an oddball stimulus in a sequence of standard stimuli evokes a negative potential as measured with Electroencephalography (EEG), which is known as the mismatch negativity (MMN) potential. The MMN response is observed in all sensory domains (Akatsuka et al., 2007; Baldeweg, 2006; Cammann, 1990; Pazo-Alvarez et al., 2003; Stagg et al., 2004) and can be interpreted under the PC framework (Garrido et al., 2009). The predictive model of the environment is updated by adjusting the brain connectivity through synaptic plasticity and dendritic spines formation upon repeated exposure of the stimuli. After a while that the standard stimulus is presented, it is reliably predicted so there is no error signalling (i.e. the MMN), which is triggered again when a deviant stimulus is presented and this pattern is reflected in the neural adjustments described above (Baldeweg, 2006; Friston, 2005). In the next section, there will be a discussion about the role of saliency or relevance of the US stimulus in the CS-US associability and how the PC framework can also account for stimulus-stimulus associations that are not behaviourally relevant nor biologically salient.

Role of task relevance in sensory predictions

The study of Pavlovian conditioning has a long tradition, and it is rooted in the animal research. Due to these circumstances, there has been always a subtle bias towards studying rewarding associations (since animals necessarily need a reward in order to perform a task) and comparatively little interest in investigating affectively neutral and task-irrelevant CS-US associations. This changed with the advent of human non-invasive neuroimaging, and researchers started to test the assumption that only rewarding associations are learned by the brain. Fletcher and colleagues (2001) investigated, using fMRI, prediction error signals regarding the associative learning of affectively neutral CS (fictitious drugs) and US (fictitious syndromes). At the beginning of the experiment, when the CS-US associability was still low, activity in the right dorsolateral prefrontal cortex (DLPFC) and the putamen was high and decreased as the associability increased. Furthermore, DLPFC activity was increasing when unexpected outcomes were presented compared to expected outcomes. McIntosh et al. (1998) used positron emission tomography (PET) to show that after a tone-light association was established, the presentation of the tone elicited activity in the visual cortex. In another study, Kok et al. (2017) studied the neural responses to CS presentation using neutrally stimuli as tones and Gabor patches using Magnetoencephalography (MEG). Participants performed an orientation detection task while the audio stimuli predicted the orientation of the Gabor stimuli. They found that indeed the tone-orientation association was learned and also that, once the association was established, the tone elicited a pre-activation of a "stimulus template". In other words, the solely presence of the CS induced a similar activation pattern of the visual stimuli in the occipital cortex. In all these studies, even if the used stimuli were not biologically relevant for the participants such as food, pain or money, the predictions made by the participants upon the stimuli are still relevant for them

because they were asked to perform these tasks. Therefore, the stimuli acquired a behavioral salience even if they do not possess any intrinsic property that made them valuable for the participants. Up to date, in the literature there are very few studies that implemented a Pavlovian conditioning procedure with truly task-irrelevant and affectively neutral stimuli. In one of these studies, researchers investigated the expectation suppression of unattended and irrelevant Gabor stimuli using fMRI (St. John-Saaltink et al., 2015). They found that, under some circumstances dependent from the manipulation of working memory, the expectation suppression was visible in retinotopically specific areas of early visual cortex (V1-V3). In another representative study, den Ouden et al. (2009) found that task-irrelevant audio-visual sensory predictions were implicitly learned by participants using fMRI, as denoted by the modulation of visual areas elicited by the predictive audio stimulus. In particular, visual cortex was progressively less activated by the predicted visual stimulus as the audio-visual association was learned. Also, expectation violations, like the absence of the predicted stimulus, triggered a gradually larger response as associative learning progressed. A possible theoretical interpretation of the reason why the brain encodes prediction errors even for irrelevant associations can be derived from the PC framework (Clark, 2013; Rao & Ballard, 1999). According to the free-energy principle, the minimization of surprise, the general goal of any living system, can be viewed as a supra goal for biological systems (Friston et al., 2006), therefore updating their internal models of the environment in order to predict potentially surprising events is also a relevant task itself. In other words, the brain is constantly trying to predict the causes of the sensory data it receives because this is, evolutionary, the best strategy one can dopt in order to control the environment and therefore having more chances to survive. Thus, making correct predictions is rewarding on itself for the brain even if these predictions are not served for behavioral relevant goals. Also, this strategy may

lead to discover new patterns of associations between stimuli that seemed irrelevant at first glance but when the environment or the goals change, these learned associations can be revaluated. In the next chapters, it will follow a series of studies that investigate the role of task relevance in sensory predictions using a combination of neuroimaging techniques, such as EEG and MEG, and computational modelling and machine learning algorithm.

Chapter 2: Stimulus-independent visual cortex activity induced by implicit auditory conditioning

Introduction

Neural systems need to continuously extract statistical regularities from the environment and update predictions about their current context to optimize behaviour (Friston et al., 2006). Traditionally, in the neuroscientific and psychological literature, prediction has been studied almost exclusively in the context of classical and instrumental conditioning paradigm (Pavlov, 1927; Skinner, 1938), which measure how living systems are able to associate neutral events (or actions) with affectively and biological significant events such as food delivery or sleep deprivation. However, it has been poorly investigated whether learning of incidental stimulus-stimulus associations (i.e., learning of associations that are irrelevant for goal-directed behaviour) is characterized by the same neuronal mechanisms of Pavlovian conditioning. The assumption in associative learning research that the strength of the association is determined by the effective salience, defined as the intrinsic property of a stimulus to elicit a biological response in the living organism, of the unconditioned stimulus (and also by the temporal contiguity between the conditioned and unconditioned stimuli) or even that the effective salience of the unconditioned stimulus is a conditio sine qua non for associative learning taking place, can be traced back to the early studies of classical and instrumental conditioning (Domjan, 2005; Treviño, 2016; Eelen, 2018). One reason for this was that the only method to study cognitive phenomena was to measure observable behavior. So, associative learning was studied exclusively with experimental designs that emphasize a behavioural response, therefore precluding the investigation of irrelevant associations that, by definition, do not exhibit a clear behavioural response. In recent times, thanks to the advent of modern neuroimaging methods, some

studies investigated the brain responses related to incidental associations, showing evidence of learning-related modulation of brain activity elicited by task-irrelevant stimuli (den Ouden et al., 2009; St. John-Saaltink et al., 2015). For instance, den Ouden et al. (2009) found that task-irrelevant audio-visual sensory predictions were implicitly learned by participants using functional magnetic resonance imaging (fMRI), as denoted by the modulation of visual areas elicited by the predictive audio stimulus. In particular, visual cortex was progressively less activated by the predicted visual stimulus as the audio-visual association was learned. Also, expectation violations, like the absence of the predicted stimulus, triggered a gradually larger response as associative learning progressed. These results can be interpreted under the general framework of predictive coding (Rao & Ballard, 1999; Friston, 2005; Clark, 2013). Predictive coding asserts that the brain is constantly using generative causal models of the world to predict and infer the causes underlying incoming sensory data in order to minimize surprise (Friston et al., 2006). These models are continuously updated using the difference between their prediction and the sensory input, a quantity that is generally referred as prediction error (Bayer & Glimcher, 2005; den Ouden et al., 2010; Schultz, 2016). These prediction errors are encoded in the neural dynamics mostly in the form of increased activity, as extensively reported in the dopaminergic system but also in other cortical and sub-cortical regions, or changes in the functional connectivity between brain areas (Schultz et al., 1997; Mehta, 2001). This kind of encoding is efficient and motivated by information-theoretic principles in the sense that it reduces redundancy by signaling only the changes rather than constancy. Here, we used Electroencephalography (EEG) to find evidence that participants learned task-irrelevant associations by solely analyzing their brain activity, without a behavioral response. Specifically, we investigated whether these learning-related patterns of brain activity elicited by implicit associations can be

explained by the same predictive coding principles governing the neural activity of associative learning for goal-directed purposes. In our experimental design, participants performed a perceptual detection task while being exposed to audio-visual distractor stimuli. Auditory distractors predicted subsequent visual distractors according to a predefined probability structure that was unknown to the participants, using a trace conditioning paradigm in which auditory stimuli preceded visual stimuli and were not temporally overlapped. Importantly, our design allows us to separate the time in which the brain generates a prediction about the next visual outcome and the time in which evaluates this prediction with the sensory data and computes the prediction error. Thus, we studied the brain responses associated with both the predictive audio stimuli (prediction) and the predicted visual stimuli (prediction error).

Methods

Participants

Twenty-one volunteers (13 females, mean age 24.3, range 19-32) participated in this study. All were right-handed with normal or corrected-to-normal vision and normal hearing, had no history of neurological disorders and were not taking any neurological medications. All participants gave informed written consent. The study was conducted in accordance with the Declaration of Helsinki and approved by the University of Trento Ethics Committee.

Procedure

During the experiment, participants were exposed to a stream of audio and visual stimuli while sitting in a dimly-lit booth at a distance of 1 m from the

CRT monitor (22.5 inch VIEWPixx; resolution: 1024 × 768 pixels; refresh rate: 100 Hz; screen width: 50 cm). Auditory stimuli consisted of low and high frequency pure tones, respectively of 250 Hz and 500 Hz. Visual stimuli consisted of Gabor patches (Fig. 1A, 4.4° × 3.4° visual angle) with Gaussian envelope, standard deviation of 18.0 and a spatial frequency of 0.08 cycles/pixel displayed in a grey background (RGB: 128, 128, 128), one with 45° orientation (right) and the other one with 135° orientation (left). On each trial, auditory stimuli predicted the presence or absence of visual stimuli according to the probability structure illustrated in Fig. 1B. One of the 2 tones (A1) was paired with one of the 2 Gabors (V1) with a probability of 90% (V1|A1), while in the remaining 10% of the times, A1 was followed by the absence of the visual stimulation (V0|A1). The other pair of stimuli (A2 and V2) were associated with an opposite statistical pattern (V2|A2 10%, V0|A2 90%). The assignment of the stimuli to the conditions was counterbalanced across the participants. The trial structure, illustrated in Fig. 1C, consisted of a fixation cross indicating the start of the trial with a duration of 100 ms, followed after 500 ms by the equally probable presentation of one of the 2 tones with a duration of 600 ms. Immediately after the offset of the audio stimulation, one of the 2 Gabors or their absence was presented for 500 ms and then the trial terminated with an inter-trial interval (ITI) of 2500 ms ± 500 ms. Importantly, this experimental design that resembles a trace conditioning paradigm (Cole et al., 1995), allows us to investigate the brain response associated to both audio (predictive) and visual (predicted) stimuli. The experiment consisted of a total of 400 trials divided in 10 blocks and the total duration of the experiment was approximately 40 minutes. Critically, in order to ensure a constant level of attention on the task and to make the statistical associations between stimuli task-irrelevant, we ask participants to perform an audio-visual target detection task. The task consisted of pressing a button whenever they perceived one of the two perceptual target

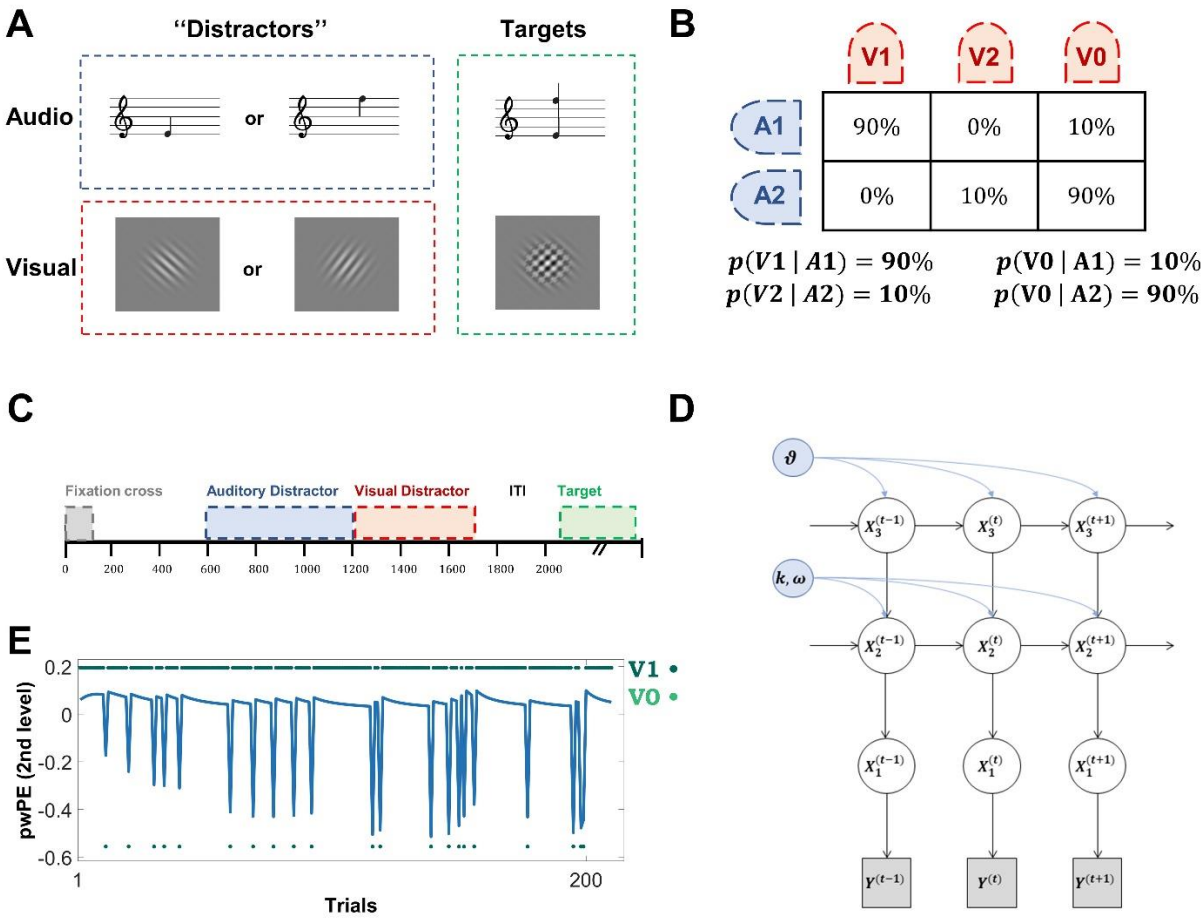

FIGURE 1 A. Stimuli presented during the experiment. The associations between ''distractor'' stimuli are those investigated in this study, the target stimuli were used to make task-irrelevant the distractors. **B.** Contingency table showing the percentage of occurrence of each visual outcome given an auditory stimulus. Below the are the conditional probabilities of the four types of trials resulting from the probability structure expressed in the table. **C.** Description of the trial structure. **D.** A graphical description of the Hierarchical Gaussian Filter (HGF) model. **E.** An example precision-weighted prediction error trajectory from one subject for the V1,V0|A1 condition.

(Fig. 1A, an auditory one that was the combination of A1 and A2, and a visual one that was the combination of V1 and V2) that was presented for 500 ms. On each block, there were 4 audio and 4 visual targets randomly presented during trial intervals and followed by an ITI. Crucially, when debriefed at the end of the experiment with a questionnaire, participants did not become aware of the

statistical associations between the stimuli. The experimental script was generated using OpenSesame with PsychoPy as backend (Mathôt et al., 2012).

EEG acquisition and preprocessing

EEG data were recorded from a standard 10-5 system with 27 Ag/AgCl electrodes cap (EasyCap, Brain Products, Germany) at a sampling rate of 1 kHz. Impedance was kept below 10 kΩ for all channels. AFz was used as the ground and the right mastoid was used as reference. Electrodes were positioned at the following scalp sites: Fpz, Fz, F3, F4, F7, F8, F9, F10, FC5, FC6, T7, C3, Cz, C4, T8, CP5, CP6, P7, P3, Pz, P4, P8, PO7, PO8, O1, Oz, and O2. All preprocessing steps were conducted using EEGLAB (Delorme and Makeig, 2004). Spherical interpolation was carried out on individual bad channels with the criterion that a channel correlated less than 0.85 on average respect to its neighbours and with the assistance of visual inspection (average number of interpolated channels: 0.74, range: 0-3). Data were down-sampled to 250 Hz and filtered with a high-pass at 0.1 Hz and a low-pass at 80 Hz, using a butterworth IIR filter with model order 2. CleanLine (Mullen, 2012) with default parameters was used to remove line noise at 50 Hz and its harmonics up to 200 Hz. After this step, data were rereferenced to a common average reference and epoched between −300 ms and 1600 ms relative to the onset of the audio stimulus with a baseline correction between -300 ms to 0 ms. Artifact rejection was applied using visual inspection and by automatically eliminating epochs containing a channel with extreme values with a threshold of ±500. The average number of trials rejected per participant was 1.1% (SD=2.1%, range 0-7.3%). Stereotyped artifacts, including blinks, eye movements and muscle artifacts were deleted via independent component analysis (ICA) using the extended infomax algorithm (Bell & Sejnowski, 1995). The average number of independent components removed was 9.33 (±3.48 SD), using a rejection

strategy based on ICLabel (Pion-Tonachini et al., 2019) and visual inspection. Finally, data were converted to Fieldtrip (Oostenveld et al., 2011) format for subsequent analyses.

EEG analysis

Data analysis aimed to assess implicit associative learning of perceptual stimuli and specifically to investigate the neural response evoked by the specific auditory stimuli repeatedly paired with the visual stimuli. To this aim, the analysis focused on two main time windows: a first epoch from 0 ms to 600 ms related to auditory stimulus presentation only, either followed or not by the Gabor patch, and a second epoch from 600 ms to 1400 ms characterized by the presentation of the Gabor patch only or by its omission. First, conventional event-related potential (ERP) analysis was performed by simply half-way splitting the data of each condition (first half vs. second half of the experiment) on the four stimulus pairs (V1|A1, V0|A1, V2|A2, V0|A2). EEG data were averaged across selected areas using frontal (Fpz, Fz, F3, F4, F7, F8, F9, F10), temporo-parietal (FC5, FC6, T7, C3, Cz, C4, T8, CP5, CP6, P7, P3, Pz), and occipital (P4, P8, PO7, PO8, O1, Oz, O2) regions of interest (ROI) (Stokes et al., 2014). Second, a regression-based approach (Myers et al., 2014; Stokes et al., 2014) aimed at assessing learning-related EEG changes on a trial-by-trial basis during exposition to audio-visual stimuli. The regression analysis was based on estimated beta values (slope parameter) obtained from a general linear model (GLM) that used the averaged EEG activity over each ROI and timepoints across block of trials as dependent variable and the number of blocks as regressor to have a proxy for the passing of time. These extracted beta parameters can be interpreted as the trend of the data over the course of the experiment. For instance, a positive slope indicates that the amplitude of the EEG signal in that particular ROI-timestep pair increased positively over time.

Trials were averaged across 10 blocks (20 trials per block) to increase the signal-to-noise ratio. This analysis was performed on the subject level for A1 and A2 time-series separately. Statistical significance was based on dependent-samples t-test (α=0.05) using mass univariate cluster-based permutation tests and maxsum as cluster statistic (Maris & Oostenveld, 2007). Third, a regression multivariate pattern analysis (MVPA), using MVPA-Light toolbox (Treder, 2020) and custom MATLAB scripts, was employed to decode differences of the activation pattern between A1 and A2 as well as between V1,V0|A1 and V2,V0|A2 (Cichy & Pantazis, 2017), considering the EEG activity in each channel averaged over blocks of trials (1-10) for each subject as features and the block order as outcome variable. A Kernel Ridge Regression (KRR, (He et al., 2014) model with the radial basis function kernel, which is a kernelized version of the ridge regression (linear least squares with L2-norm regularization) that allows non-linear mappings of the data, was applied over time. Z-scoring was applied across samples for each time point separately to normalize channel variances and remove baseline shifts. Model performance was estimated using the Root Mean Square Error (RMSE) as metric and repeated k-fold cross-validation with 5 repetitions and 5 folds, to avoid overfitting and increase robustness of results. Cluster-based permutation tests with the same hyperparameters of the previous regression analysis were considered for estimating statistical significance. In addition, searchlight analysis (Kriegeskorte et al., 2006) applied to each channel assessed spatial features relevance. Analysis of V1 and V2 stimuli was also based on regression analysis and MVPA but in order to consider the intrinsic differences of their probability distributions were applied to additional variables (see next section) estimated using the Hierarchical Gaussian Filter (HGF) model (Mathys et al., 2014).

Hierarchical Gaussian Filter model

The Hierarchical Gaussian Filter model (HGF) is a Bayesian generative model (Mathys et al., 2014) of perceptual inference on a changing environment based on sequential input (Iglesias et al., 2013; Hauser et al., 2014; Powers et al., 2017). The HGF consists of perceptual and response models, representing a Bayesian observer who receives a sequence of inputs, updates an internal model of how environment generates those inputs and predicts future observations (Fig. 1 D). Since our experimental design deliberately precluded behavioral responses, we used only the perceptual model (Stefanics et al., 2018). In such a modeling framework, a perceptual model comprises a hierarchy of 3 hidden states (x), which account for a multi-level belief updating process about the hierarchically related environmental states giving rise to sensory inputs, and an observation model (y) represents the actual occurrence of a stimulus in a given trial (Fig. 1). The model assumes that environmental hidden states evolve conditionally on the states at the immediately higher level. The hidden states process at the first level of the perceptual model represents a sequence of beliefs ($x_1^{(t)}$) about stim-ulus occurrence, that is, whether a visual stimulus was present ($y^{(t)} = 1$) or absent ($y^{(t)} = 0$) at trial t, and is modelled as follows:

$$x_1^{(t)} | x_2^{(t)} \sim Bernoulli(s(x_2^{(t)})) \quad (1)$$

where $s(x_2^{(t)}) := [1 + \exp(x_2^{(t)})]^{-1}$ is the logistic sigmoid function. Here, the hidden states at the second level ($x_2^{(t)}$) is an unbounded real parameter of the probability that $x_1^{(t)} = 1$, thus representing the current belief of the probability that a given stimulus occurs. Such an hidden state process evolves according to a Gaussian random walk:

$$x_2^{(t)}|x_2^{(t-1)}, x_3^{(t)} \sim Gaussian(x_2^{(t-1)}, \exp(\kappa x_3^{(t)} + \omega)) \quad (2)$$

which depends on both its value at a previous trial t, and the hidden state at the third level of the hierarchy. In particular, the higher-level hidden state process ($x_3^{(t)}$) determines the log-volatility of the hidden state process at the second level, thus codifying the volatility of the environment during the time course of the experiment. This process evolves according to a Gaussian random walk:

$$x_3^{(t)}|x_3^{(t-1)} \sim Gaussian(x_3^{(t-1)}, \vartheta) \quad (3)$$

The parameter set $(\kappa, \omega, \vartheta)$ determines the dispersion of the random walks atdifferent levels of the hierarchy and allows to shape individual difference in learning. By inverting the generative model, given a sequence of observations (y), it is possible to obtain the updating process of the trial-by-trial estimates of the hidden state variables. The update rules share a common structure across the model's hierarchy: at any level i the update of the posterior mean $\mu_i^{(t)}$of the state x_i, that represents the belief on trial k, is proportional to the precision-weighted prediction error (pwPE) $\varepsilon_i^{(t)}$ as follows:

$$\mu_i^{(t-1)} - \mu_i^{(t)} \propto \psi_i^{(t)} \delta_{i-1}^{(t)} = \varepsilon_i^{(t)} \quad (4)$$

$$\psi_i^{(t)} = \frac{\hat{\pi}_{i-1}^{(t)}}{\pi_i^{(t)}} \quad (5)$$

$$\pi_i^{(t)} = \frac{1}{\sigma_i^{(t)}} \quad (6)$$

As shown in Eqs 3–5, in each trial, a belief update $\mu_i^{(t-1)} - \mu_i^{(t)}$ is proportional to the prediction error at the level below $\delta_{i-1}^{(t)}$. The pwPE is the product of the

prediction error $\delta_{i-1}^{(t)}$ and a precision ratio $\psi_i^{(t)}$ that depends on the precision (inverse variance, Eq. 5) of the prediction at the level below $\hat{\pi}_{i-1}^{(t)}$ and the current level $\pi_i^{(t)}$. In this application, we are interested in the update equations of the hidden states at the second level, which have a general form similar to those of traditional reinforcement learning models, such as the Rescorla-Wagner model (Rescorla & Wagner, 1972). The pwPE on the second level, is thus assumed to be responsible for the learned perceptual association. The nature of the pwPE can be described through the following update equation of the mean of the second level:

$$\mu_2^{(t)} = \mu_2^{(t-1)} + \sigma_2^{(t)}(\mu_1^{(t)} - s(\mu_2^{(t-1)})) \quad (4)$$

where the last term represents the prediction error $(\mu_1^{(t)} - s(\mu_2^{(t-1)}))$ at the first level weighted by the precision term $\sigma_2^{(t)}$ (see (Mathys et al., 2014) for a general derivation and more mathematical details). Trajectories of pwPEs with separate models for A1 and A2 were calculated by estimating the parameters that minimize Bayesian Surprise using the Broyden-Fletcher-Goldfarb-Shannon (BFGS) quasi-Newton optimization algorithm. We determined these Bayes optimal perceptual parameters by inverting the perceptual model based on the stimulus sequence alone and a predefined prior for each parameter (the standard in the HGF toolbox, version 5.2 implemented via the Translational Algorithms for Psychiatry Advancing Science toolbox). These model-derived trajectories of pwPEs from the second level were used (Fig. 1 E) as, respectively, regressor and outcome variables in the GLM-based regression and MVPA.

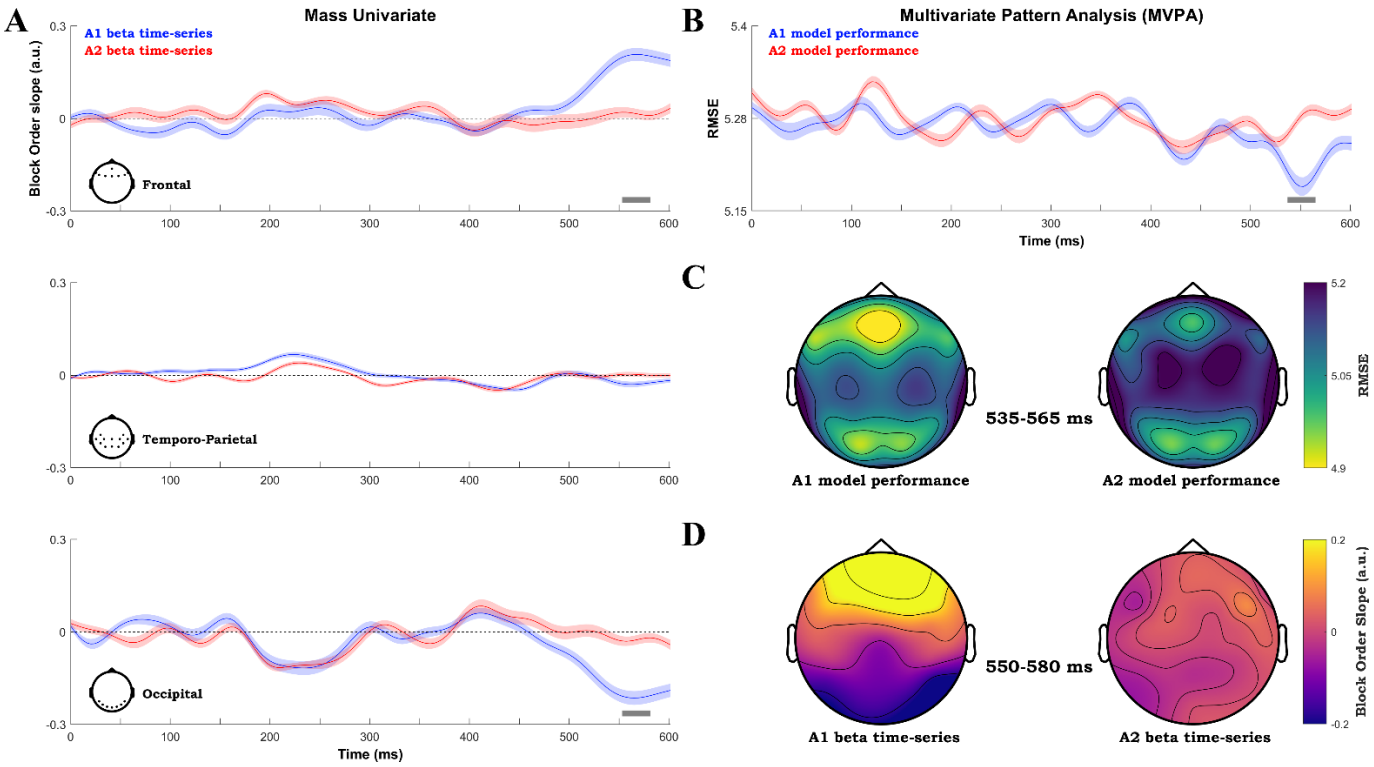

FIGURE 2 A. Regression slope time series estimated across 10 blocks at each timestep in Frontal, Temporo-Parietal and Occipital ROIs for A1 (in blue) and A2 (in red) conditions. Shades indicate standard error of the mean (SEM). Horizontal bars represent significant differences between the two conditions with $p < 0.05$ (cluster-corrected). **B.** MVPA performance for A1 (in blue) and A2 (in red) across time using all channels as features. Shading indicates SEM across subjects and the horizontal bar shows a significant difference between the two models ($p < 0.05$ cluster-corrected). **C.** Topographical representation of the Searchlight analysis at the latency resulted significant in the MVPA analysis across time. **D.** Topographical representation of the results of the regression analysis at the latency resulted significant.

Results

Participants debriefed at the end of the experiment reported not to be consciously aware of audio-visual stimuli pairings. They reported to have noticed neither any particular regularity of stimuli presentation nor any pairing between auditory and visual stimuli when specifically interrogated on possible audio-visual associations.

Conditioned stimuli

ERP analysis of the conditioned auditory stimuli (1st epoch: 0-600ms) showed a significant difference between the A1 first and second half waveform in the frontal and occipital ROI around 520 and 580 ms (Fig. 3 A, $p<0.05$, cluster corrected), while for the comparison of A2 first and second half ERPs no significant differences emerged (Fig. 3 B). Interestingly, the occipital region in A1 in the second part of the experiment was decreasing its activity with respect to the first part, while for the frontal region we observed the opposite pattern. Regression analysis of beta values revealed a significant difference ($p < 0.05$, cluster corrected) between A1 and A2 in the frontal and occipital ROIs; specifically, A1 with respect to A2 showed increased positivity in the frontal ROI and increased negativity in the occipital ROI in a time window immediately preceding V1 presentation, between 550 and 580 ms in the frontal and occipital areas (Fig. 2 A-D). MVPA analysis over time showed significantly higher performance of the kernel ridge regression model for A1 as compared to A2, in a time window between 535 and 565ms, as evidenced by the lower residual variance expressed through the RMSE. Searchlight analysis indicated a major contribution of frontal and occipital ROIs to the observed differences in the regression model between A1 and A2 (Fig. 2 B-C).

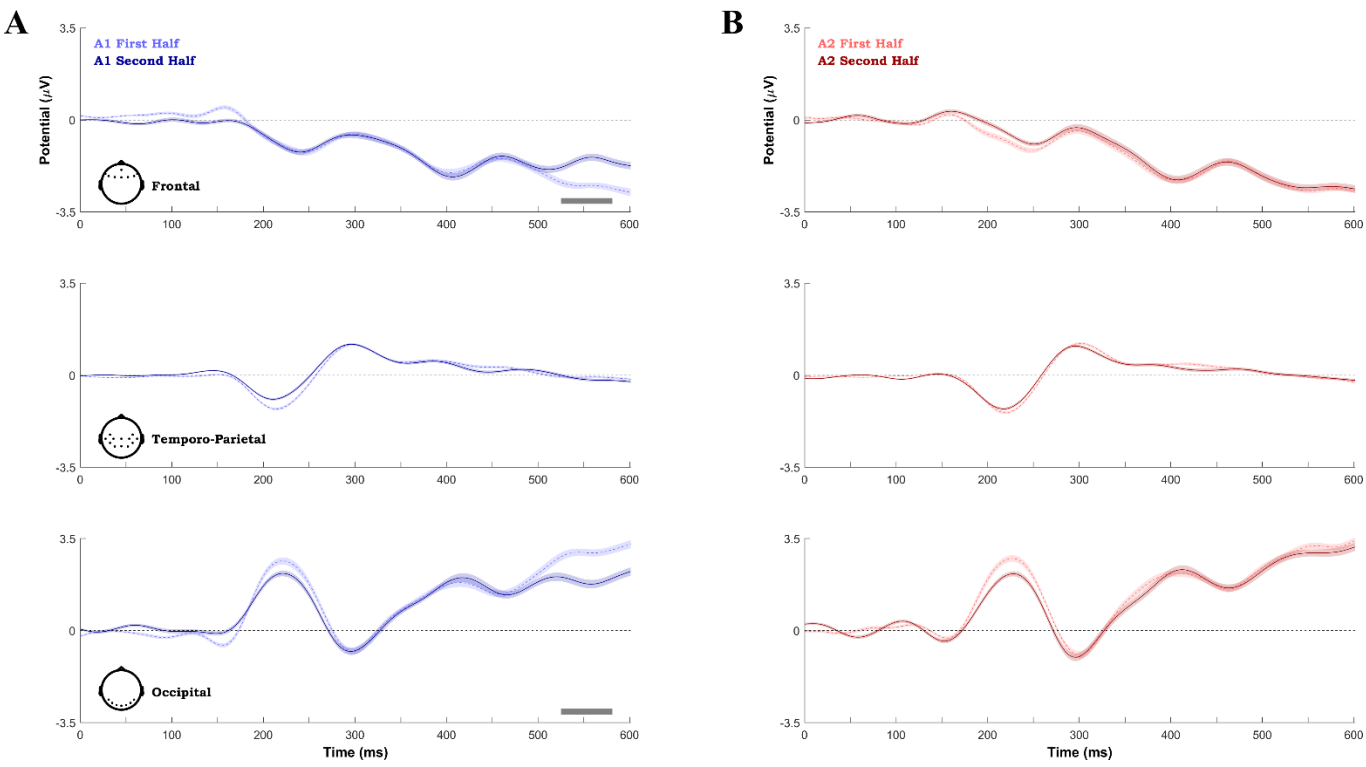

FIGURE 3 A. ERPs computed on first (in dashed light blue) and second (in solid dark blue) half of trials of A1 condition across the frontal, temporo-parietal and occipital ROIs. Shading indicates SEM. **B.** ERPs computed on first (in dashed light red) and second (in solid dark red) half of trials of A2 condition across the frontal, temporo-parietal and occipital ROIs. Shading indicates SEM.

Unconditioned stimuli

ERP analysis of the unconditioned stimuli (2nd epoch: 0-600 ms after unconditioned stimuli onset) showed no significant effects (Fig. 5 A-B) in the selected ROIs comparing first and second half of all visual conditions (V1|A1, V0|A1, V2|A2, V0|A2). Regression analysis of pwPE trajectories - estimated by considering the opposite probability distribution of V1 and V0 occurrence given A1 with respect to V2 and V0 occurrence given A2, showed a significant difference of EEG pattern between V1+V0 and V2+V0 in the time window 240-280 ms after the offset of the conditioned stimulus ($p<0.05$, cluster corrected) in the occipital ROI and around 240-300 ms in the frontal ROI (Fig. 4, A-D;

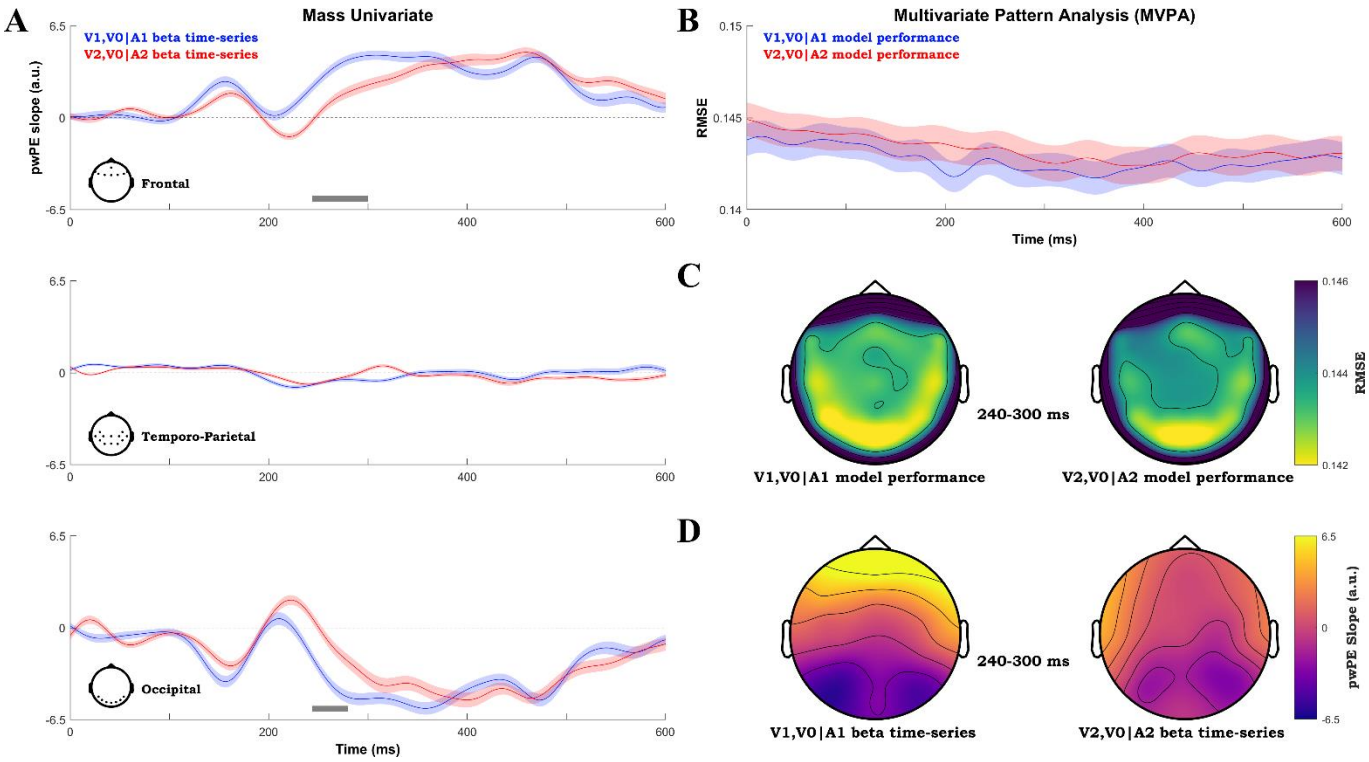

FIGURE 4 A. Regression slope time series estimated on single trial at each timestep in Frontal, Temporo-Parietal and Occipital ROIs, using the pwPE trajectories estimated from the HGF model for A1 (in blue) and A2 (in red) conditions. Shades indicate SEM and horizontal bars represent significant differences between the two conditions ($p < 0.05$ cluster-corrected). **B.** MVPA performance for A1 (in blue) and A2 (in red) across time using all channels as features. Shading indicates SEM across subjects and the horizontal bar shows a significant difference between the two models ($p < 0.05$ cluster-corrected). **C.** Topographical representation of the Searchlight analysis at the latency resulted significant in the MVPA analysis across time. **D.** Topographical representation of the results of the regression analysis at the latency resulted significant.

p<0.05, cluster corrected). MVPA over time did not evidenced a significant difference of KRR model performance between V1,V0|A1and V2,V0|A2(Fig. 4B). Searchlight analysis, performed in the same time window resulted significant in the GLM analysis (240-300 ms), revealed a greater contribution of temporal-occipital regions for predicting pwPE trajectories in V1,V0|A1 model performance, while for V2,V0|A2 only a restricted number of occipital channels were contributing most (Fig. 4C).

Discussion

In this study, we investigated the neural mechanisms underlying task-irrelevant sensory prediction and prediction errors. We presented to the participants audio-visual associations without requiring a behavioural response. Instead, they performed a perceptual detection task concurrently, thus intentionally directing their attention away from the audio-visual associations and making them irrelevant for the task they were instructed to perform. Task-irrelevant stimuli were presented using a trace conditioning approach, allowing us to study the separate contribution of sensory prediction and prediction errors to the modulation of the brain activity. We found that participants learned these incidental associations without being aware, as none of them reported awareness about the audio-visual association in the debriefing questionairre, by analyzing their brain responses. One key finding of this study was related to the brain activity associated to the audio predictive stimuli. We found a significant modulation of frontal and occpital activity over time triggererd by the presence of the audio stimulus that was more predictive. Specifically, frontal activity was progressevely more positive while occipital activity was more negative as the audio stimulus gained predictive power. Both these trends were observed around 100 ms before the appearance of the visual outcome. These results evidenced that incidental stimulus-stimulus associations elicited a preparatory activity in the brain, a sign that the sensory prediction took place. Also, a similar pattern was observed in task-relevant audio-visual associations in McIntosh et al. (1998), in which they showed that the auditory stimulus was able to evoke responses in the visual cortex, and in Kok et al. (2017), in which they were able to decode the orientation of visual grating stimuli based on a preceding auditory stimulus. This proactive, anticipatory activity is in line with predictive coding theories because it allows sensory cortices to be prepared for upcoming sensory data by efficiently preallocating neural resources (Kok et al., 2017). Moreover,

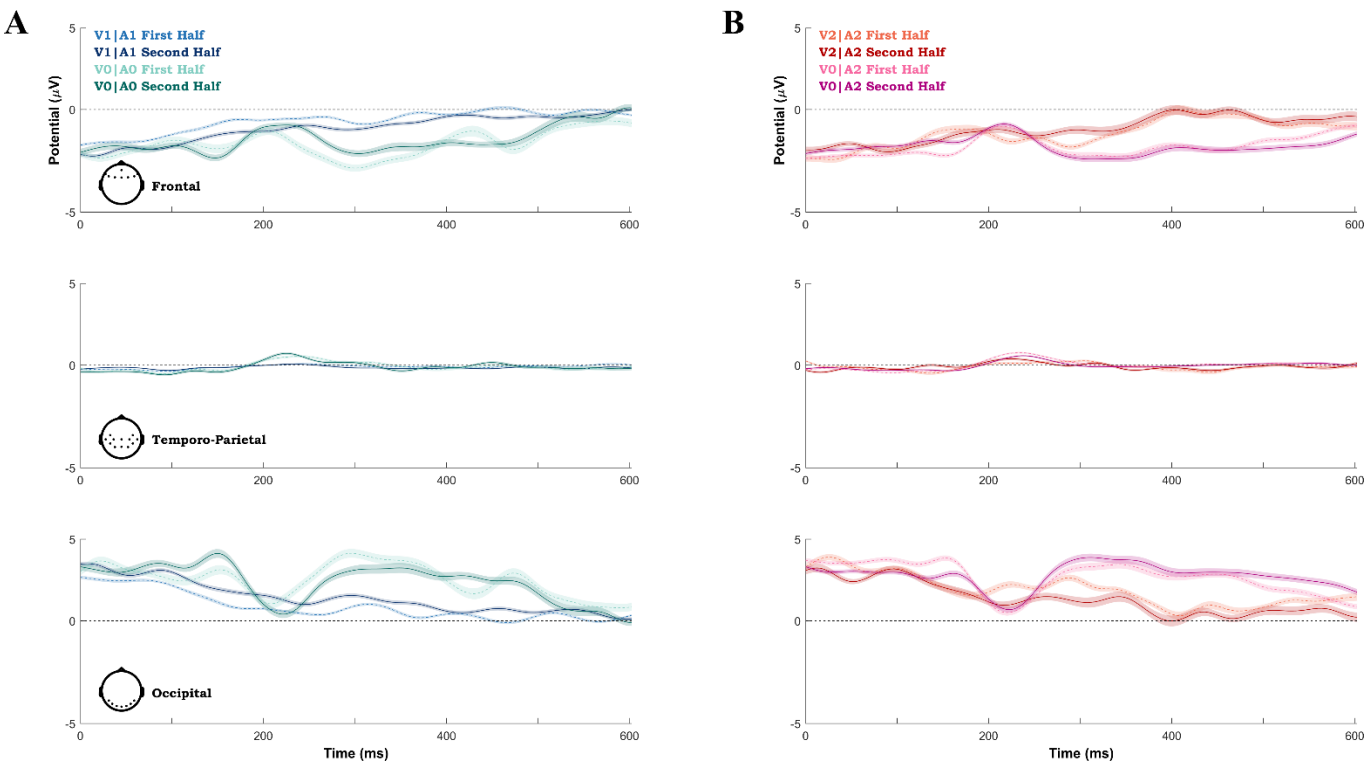

FIGURE 5 A. ERPs computed on first and second half of trials of A1-V1 (first in dashed light blue, second in solid dark blue) and A1-V0 (first in dashed light green, second in solid dark green) conditions across the frontal, temporo-parietal and occipital ROIs. Shading indicates SEM. **B.** ERPs computed on first and second half of trials of A2-V2 (first in dashed light red, second in solid dark red) and A2-V0 (first in dashed light pink, second in solid dark pink) conditions across the frontal, temporo-parietal and occipital ROIs. Shading indicates SEM.

the anticipatory activity found in this study is relevant not only because it extends the evidence for this phenomenon to task-irrelevant settings, but also because the anticipatory effects of predictive processes are less explored in general, since most of the studies in this field focused their attention on prediction error signals or a mixture of prediction and prediction-error signals to study how predictive coding principles are implemented in the brain (Dürschmid et al., 2019). Another key finding of this study was that visual activity was modulated by prediction error signals only when preceded by predictive stimuli. Specifically, brain activity in the occipital ROI triggered by

the visual outcome covaried with pwPE trajectories extracted from the HGF model only when preceded by the predictive audio stimulus, as shown by the GLM analysis. These results are in line with the findings of den Ouden et al. (2009) that showed the same pattern of activity in the visual cortex for task-irrelevant stimuli using fMRI. Since we used EEG, we can extend these results by characterizing the timing of this prediction error related modulation. Thus, we can conclude that the prediction error related modulation of the occipital activity started about 200-300 ms after the onset of the visual outcome. These findings demonstrate prediction error signals are computed even for irrelevant predictions and implemented similarly as in relevant predictions. One could speculate that this happens because the minimization of surprise can be viewed as a supra goal for biological systems (Friston et al., 2006), therefore updating their internal models of the environment in order to predict potentially surprising events is also relevant (den Ouden et al., 2009). In summary, we proved task-irrelevant sensory predictions draw upon the same neural mechanisms of behaviourally-relevant sensory associative learning. Our study advanced the notion that prediction errors are computed for task-irrelevant predictions by sheding some light about the latency of this process. Also, to the best of our knowledge, our study is the first to demonstrate an anticipatory activity for task-irrelevant predictions that was found similarly in other studies for task-relevant predictions. Moreover, we provided detailed information about the latency and spatial location of this process. A limitation of our study was that we used a model-based approach only for modeling prediction error signals and that we limited our investigation to scalp activity. A possible extension would be to extract different metrics from the computational model related to predictive processing phenomena, such as the Bayesian Surprise (Itti & Baldi, 2009) that can be seen as a proxy for the model update process (O'Reilly et al., 2013). Future extensions of our current work may also include computational

models able to simulate both prediction and prediction error trial-wise trajectories (eg. The Temporal Difference Model) and time-frequency and connectivity analysis to further investigate the frequency bands responsible for propagating these predictive signals throughout the sensory cortices and how the strength of the connections between sensory areas are modulated by predictive processes.

Chapter 3: Task-irrelevant sensory associations modulate visual oscillatory activity in the beta band

Introduction

Associative learning is a fundamental ability that biological systems possess in order to adapt to a nonstationary environment. This phenomenon has been studied extensively in the last century, but in the recent period there has been some major breakthroughs in the theoretical framework that attempts to capture its essential features (Delamater & Matthew Lattal, 2014). One of these paradigm shifts regards the notion that the central neural system generates internal predictions to anticipate the causes of its perceptual experience and compute a prediction error to update its generative model of the environment, an idea generally known as the predictive coding framework (Rao & Ballard, 1999; Friston, 2010; Clark, 2013). However, it is not clear whether the brain generates these predictions only for goal-oriented behavior or they are more a general characteristic of the brain function. This lack of knowledge about irrelevant stimulus-stimulus associations can be explained by the fact that associative learning has been almost exclusively studied in the past with animals, which posits serious difficulties for studying non rewarding associations. Here, we test the effects of task-irrelevant prediction errors on the modulation of time-frequency representations of the EEG signal from human participants. We designed a task in which participants, similarly to the first study in this thesis, performed a perceptual detection task while being exposed to audio-visual distractor stimuli. Auditory distractors predicted visual distractors according to a predefined transition probability matrix that was unknown to the participants, using a trace conditioning paradigm in which

auditory stimuli anticipated visual stimuli and were not temporally overlapped. In this study, we added two differences with respect to the first study. First, we exposed participants to two sessions called Habituation and Extinction, before and after the conditioning, in which the transition probabilities were all equal. Second, we spread the course of the experiment into a two week period, in order to test the stability of this learning effect and also strengthen it. Finally, we focused our analysis only on the unconditioned stimuli in order to test the effect of prediction errors and not predictions.

Methods

Participants

Eight volunteers (4 females, mean age 26.7, range 22-38) participated in this study. A detailed explanation about the choice of the sample size can be found in the EEG analysis section. All were right-handed with normal or corrected-to-normal vision and normal hearing, had no history of neurological disorders and were not taking any neurological medications. All participants gave informed written consent. The study was conducted in accordance with the Declaration of Helsinki and approved by the University of Trento Ethics Committee.

Procedure

The experiment is divided in 3 sessions (Habituation, Acquisition and Extinction) and was conducted in 6 days spanning across 2 weeks. In first and last days we recorded EEG data while participants were exposed to Habituation and Extinction sessions, respectively. In the four middle days the Acquisition

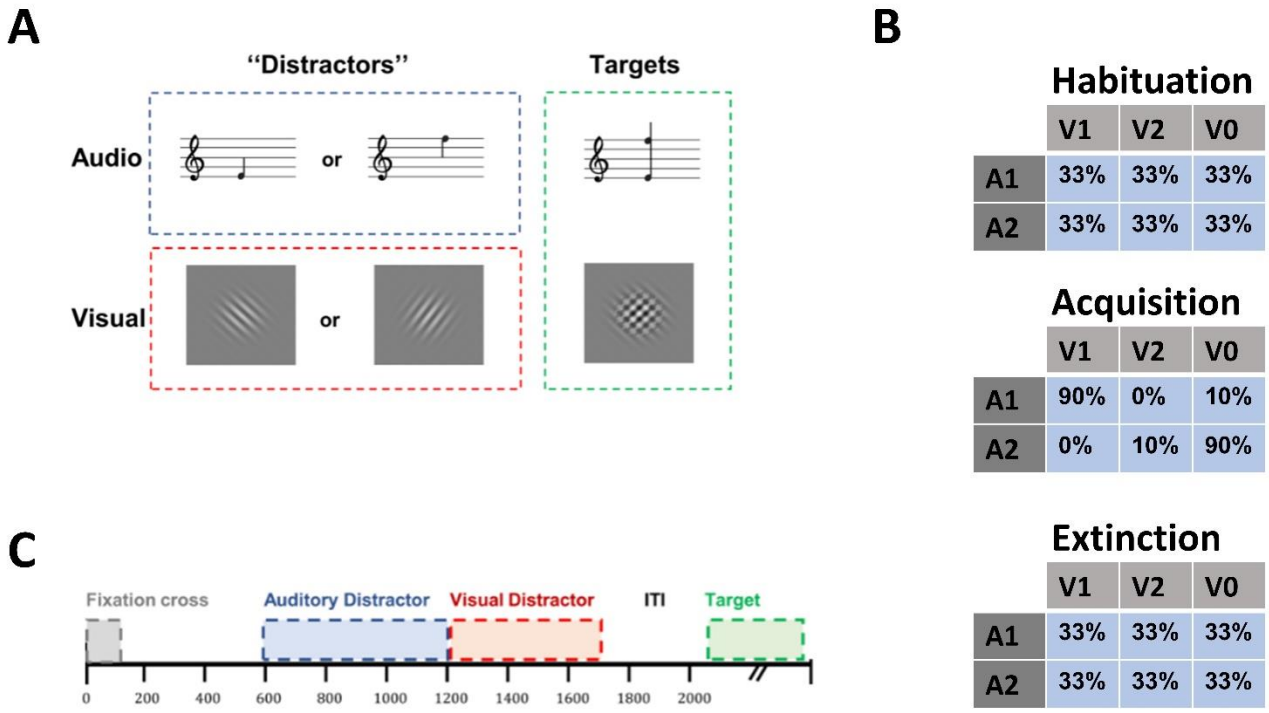

FIGURE 1 A. Stimuli presented during the experiment. The associations between ''distractor'' stimuli are those investigated in this study, the target stimuli were used to make task-irrelevant the distractors. **B.** Contingency table showing the percentage of occurrence of each visual outcome given an auditory stimulus for the three sessions Habituation, Acquisition and Extinction. **C.** Description of the trial structure.

session was run. During the all experiment, participants were exposed to a stream of audio and visual stimuli while sitting in a dimly-lit booth at a distance of 1 m from the CRT monitor (22.5 inch VIEWPixx; resolution: 1024 × 768 pixels; refresh rate: 100 Hz; screen width: 50 cm). Auditory stimuli were 2 low and high frequency tones (Fig. 1A), respectively of 250 Hz and 500 Hz. Visual stimuli were 2 Gabor patches (Fig. 1A, 4.4° × 3.4° visual angle) with Gaussian envelope, standard deviation of 18.0 and a spatial frequency of 0.08 cycles/pixel displayed in a grey background (RGB: 128, 128, 128), one with 45° orientation (right) and the other one with 135° orientation (left). On each trial (Fig. 1C), auditory stimuli predicted the presence or absence of visual stimuli according

to the probability structure illustrated in Fig. 1B. What differed across the sessions was indeed the probability structure of the stimuli association. In Acquisition, one of the 2 tones (A1) was paired with one of the 2 Gabors (V1) with a probability of 90% (A1-V1), while in the remaining 10% of the times, A1 was followed by the absence of the visual stimulation (A1-V0). The other pair of stimuli (A2 and V2) were associated with an opposite statistical pattern (A2-V2 10%, A2-V0 90%). In Habituation and Extinction, the probability that given one sound one could predict the presence of one visual stimuli was identical for all stimuli (33% for A1-V1, A1-V2, A1-V0, A2-V1, A2-V2, A2-V0). The assignment of the stimuli to the conditions was counterbalanced across the participants. The trial structure, illustrated in Fig. 1C, consisted of a fixation cross indicating the start of the trial with a duration of 100 ms, followed after 500 ms by the equally probable presentation of one of the 2 tones with a duration of 600 ms. Immediately after the offset of the audio stimulation, one of the 2 Gabors (or their absence) was presented for 500 ms and then the trial terminated with an inter-trial interval (ITI) of 2500 ms ± 500 ms. The experiment consisted of 300 trials divided in 10 blocks for the Habituation and Extinction session each, and 1200 trials for the Acquisition (300 trials divided in 10 blocks for each day). Critically, in order to ensure a constant level of attention on the task and to make the statistical associations between stimuli task-irrelevant, we ask participants to perform an audio-visual target detection task. The task consisted of pressing a button whenever they perceived one of the two perceptual target (Fig. 1A, the auditory target was the combination of A1 and A2, and the visual target was the combination of V1 and V2) that was presented for 500 ms. On each block, there were 4 audio and 4 visual targets randomly presented during trial intervals and followed by an ITI. Crucially, when debriefed at the end of the experiment with a questionnaire, participants did not become aware of the

statistical associations between the stimuli. The experimental script was generated using OpenSesame with PsychoPy as backend (Mathôt et al., 2012).

EEG acquisition and preprocessing

EEG data were recorded from a standard 10-5 system with 64 Ag/AgCl electrodes cap (EasyCap, Brain Products, Germany) at a sampling rate of 1 kHz. Impedance was kept below 10 kΩ for all channels. AFz was used as the ground and the right mastoid was used as reference. All preprocessing steps were conducted using EEGLAB (Delorme & Makeig, 2004). Spherical interpolation was carried out on individual bad channels with the criterion that a channel correlated less than 0.85 on average respect to its neighbours and with the assistance of visual inspection (average number of interpolated channels: 2.83, range: 1-5). Data were down-sampled to 250 Hz and filtered with a high-pass at 1 Hz and a low-pass at 80 Hz, using a butterworth IIR filter with model order 2. CleanLine (Mullen, 2012) with default parameters was used to remove line noise at 50 Hz and its harmonics up to 200 Hz. After this step, data were rereferenced to a common average reference and epoched between -500 ms and 1500 ms relative to the onset of the visual stimulus to avoid edge artifact for the time-frequency analysis. Artifact rejection was applied using visual inspection and by automatically eliminating epochs containing a channel with extreme values with a threshold of ±500. The average number of trials rejected per participant was 1.8% (SD=2.3%, range 0-6.5%). Stereotyped artifacts, including blinks, eye movements and muscle artifacts were deleted via independent component analysis (ICA) using the extended infomax algorithm (Bell & Sejnowski, 1995). The average number of independent components removed was 17.51 (±7.92 SD), using a rejection strategy based on ICLabel (Pion-Tonachini et al., 2019) and visual inspection. Finally, data were averaged across pre-defined frontal, temporal, parietal and occipital regions of interest

(ROI) and were converted to Fieldtrip (Oostenveld et al., 2011) format for subsequent analyses.

EEG analysis

In this study, we investigated the oscillatory activity associated with the predicted visual stimuli. For this reason, we computed the time-frequency representation of the EEG data in each epoch by convolving each ROI time-series with a set of complex Morlet wavelets and then taking the inverse Fast Fourier Transform. The wavelets were defined as $e^{i2\pi tf}e^{-t^2/(2\sigma^2)}$, where *t* is time, *f* is frequency, and σ defines the width of each frequency band, set according to *n*/(2πf), where *n* is the number of wavelet cycles. The frequency *f* increased from 8 to 45 Hz in 20 logarithmically spaced steps, and the number of cycles *n* increased from 3 to 12 in 20 logarithmically spaced steps. From the resulting complex signal, the power of each frequency at each time point was obtained. The power was baseline-normalized to decibel (dB) in respect to -400 ms and -100 ms interval relative to the onset of the audio stimulus. We re-epoched the trials from 0 ms to 800 ms to get rid of the edge artifacts. To increase the signal-to-noise ratio, trials were averaged across 10 blocks for each condition (50 trials averaged per block for 6 conditions). Finally, we subtracted the Habituation blocks from the Extinction blocks and compared the conditions having the same visual outcome (A1-V1 vs A2-V1; A1-V2 vs A2-V2; A1-V0 vs A2-V0). To perform statistical analyses on the group level and at the same time having enough samples, we concatenated all the blocks from all participants into one super-participant for each condition. This allowed us to be able to perform statistical analysis at the group level even with a small sample of individuals. For each comparison we had, therefore, 160 samples (10 blocks per participant per condition) belonging in an equal split to the two conditions to compare. Then, we performed an a priori power analysis to estimate if the

sample size was adequate using G*Power toolbox (Faul et al., 2007). We used an independent two-samples two-tails t-test as statistical test and set the alpha parameter to 0.05, the power to 0.8 and the effect size to 0.5. The result was that the required sample size was of 128, confirming that the samples we had were acceptable. In order to further increase the statistical power, we collapsed the frequency dimension into 3 bands (Alpha 8-13 Hz, Beta 13-30 Hz and Gamma 30-45 Hz) and performed statistical analyses across them. To assess statistical significance for each of the 3 comparisons, we ran for each ROI-frequency band pair mass univariate cluster-based permutation tests with an independent-samples t-test (α=0.05), 10000 permutations and maxsum as cluster statistic.

Results

Statistical analyses across all the comparisons and ROI-frequency band pairs revealed significant results only in the A1-V1 vs A2-V1 comparison over the Frontal ROI and the Beta frequency band (Fig 2, A). Specifically, we found two significant temporal clusters ($p<0.05$) in the beta band power spectrum time-series (Fig 2, B) in the 0-110 ms and 370-510 ms intervals from the onset of V1. In the first significant time segment, there was an increase across the experiment in the power spectrum relative to the trials in which the predictive stimulus (A1) was present, while for the unpredictive stimulus (A2) was not the case. Crucially, it is clear that this increased beta activity is starting even before the onset of the visual stimulus, indicating that what triggered this pattern was the presence of the predictive stimulus. In addition, from the topography we can observed that the occipital regions were also increasing their beta activity across the time course of the experiment (Fig 2, C), but this pattern did not survived to statistical testing. For the second significant cluster of time points, we observed

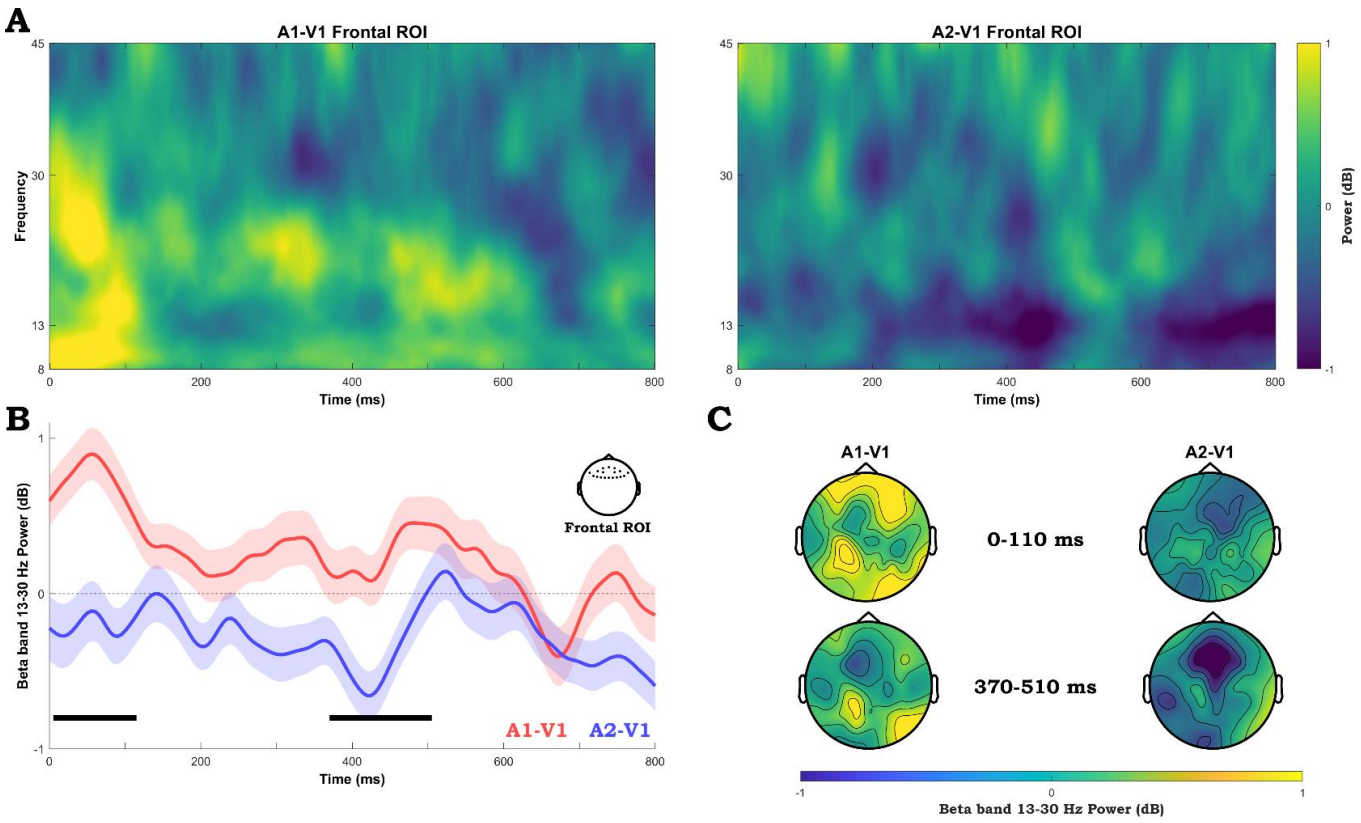

FIGURE 2 A. Time-frequency representation of A1-V1 and A2-V1 over the frontal ROI from the onset of the visual stimulus. **B.** Power spectrum time-series in the beta band for the two conditions. **C.** Topography of the statistically significant time segments for the two conditions.

that the beta activity triggered by the visual outcome preceded by A2 was decreased across the experiment, while for A1 did not change.

Discussion

In this study, we investigated time-frequency representations of the EEG signal underlying task-irrelevant associations. We presented to the participants audio-visual associations while performing a perceptual detection task, thus intentionally directing their attention away from the audio-visual associations and making them irrelevant for the task they were instructed to perform. We

found that participants learned these associations without being aware, as none of them reported knowledge about the sensory associations in the debriefing questionairre, by analyzing their brain responses. The key finding of this study was a difference in the modulation of the beta band induced by the presentation of the V1 stimulus preceded by the predictive and unpredictive cue A1 and A2. We found that beta band power increased when preceded by A1 respect to A2 even before the onset of V1. This could be in line with the findings of the first study about the anticipatory activity of the conditioned stimulus. Moreover, we found a decremented beta power after 350 ms the onset of V1 when preceded by A2 respect to A1. This result could be interpreted as a signal of prediction error since A2-V1 is an association that violates the expectation of the participants. These findings are in line with previous studies that found differences in the beta band induced by sensory predictions (van Ede et al., 2011; Todorovic et al., 2015). Taken togher, we demonstrated that task-irrelevant associations are captured by the brain even when spread across a long time range such as in this experiment. This can be interpreted within the framework of predictive coding, claiming that avoid surprising events can be viewed as a meta goal for biological agents (Friston et al., 2006), therefore updating their internal models of the environment in order to predict potentially surprising events is also relevant (den Ouden et al., 2009). In addition, the study of oscillatory activity can shed more light on the underlying mechanisms of sensory predictions, since it has been shown that different frequency bands may carry information about different stages of the predictive processing (Arnal & Giraud, 2012). The current view about the oscillatory dynamics of predictive processing is that there is an asymmetry between forward and backward passing of information (Bastos et al., 2015). Specifically, ascending prediction errors are conveyed at a faster frequency, for example gamma band, while descending predictions are encoded at lower frequencies such as alpha and beta band

(Friston, 2019). The reason why we found a lower frequency band modulation due to prediction error encoding (beta) may be related to the fact that predictive processing are always active about multiple aspects of the sensory experience. For example, if the goal of a biological system is to predict a visual input, the system have to combine different expectations about what was the cause of the visual stimulus and where that object was. Thus, the system have to use a nonlinear mixture of top-down predictions about what and where the visual input is. A limitation of this study was the low sample size, due mainly to the long duration of the experiment and therefore the high level of dropout among the participant. Future research may account for this considering a shorter period of exposition to the stimuli associations by increasing the percentage of association.

Chapter 4: Revealing the similarity of relevant and irrelevant associations induced brain dynamics

Introduction

In an ever-changing environment, being able to predict future events is a fundamental aspect of the behavior of sentient systems. It allows adaptation by enabling the system to be prepared for possible threats or opportunities. Extensive evidence suggests that the brain actively generates predictions about the causes of the gathered sensory data in order to optimize behavior (Friston, 2010), a theoretical framework known as Predictive Coding (PC; Rao & Ballard, 1999; Clark, 2013). PC advocates the brain is constantly using generative causal models of the environment to avoid surprising events (Friston et al., 2006). These models are continuously updated by the difference between their expectations and the sensory input, a quantity usually termed prediction error (Bayer & Glimcher, 2005; den Ouden et al., 2010; Schultz, 2016). These prediction errors are visible in the brain mostly in the form of increased activity or changes in the connectivity between brain regions (Schultz et al., 1997; Mehta, 2001). One key notion that biological systems face since their birth is that not all statistical regularities captured in the environment are important for predicting behaviorally relevant events. However, little attention has been paid to the investigation of relevant and irrelevant sensory predictions and how they are encoded in the brain. Den Ouden et al. (2009) found that task-irrelevant audio-visual sensory associations were implicitly learned by participants using functional magnetic resonance imaging (fMRI), as denoted by the modulation of visual areas triggered by the conditioned audio stimulus. In particular, occipital regions were progressively less activated by the predicted visual stimulus as the audio-visual association was learned. Also, expectation violations, like the absence of the predicted stimulus, elicited a larger response

as learning progressed. In another study, Stokes et al. (2014) investigated directly the difference between relevant and irrelevant associations. They exposed participants to a stream of complex fractal images and ask them to press a button when they saw one of these fractal. In their design, the target fractal was predicted by another fractal, the task-relevant association, while there were two different images that were associated with the same probability as the cue-target association, regarded as the task-irrelevant association. They found a strong modulation of the response elicited by the task-relevant association after 200 ms post-target in central and posterior electrodes, but no corresponding effects for task-irrelevant stimuli. Here, we used Magnetoencephalography (MEG) to study the brain activity and functional connectivity networks to relevant and irrelevant associations. We exposed participants to audio-visual pairings asking them to press a button whenever they perceived their target visual stimulus. We manipulated the probability structure relating auditory and visual stimuli to increase the associability between one auditory stimulus and the target stimulus, while manipulating the conditional probability between another audio-visual pair of stimuli in a symmetric way with respect to the cue-target association. Before and after this session, participants were presented the same stimuli but with an equal probability of association for all stimuli. This allowed us to invesigate the difference before and after associative learning took place in a relevant and irrelevant setting and to compare their neural effects.

Methods

Participants

Twenty volunteers (11 females, mean age 24.1, range 19-35) participated in this study. All were right-handed with normal or corrected-to-normal vision and normal hearing, had no history of neurological disorders and were not taking

any neurological medications. All participants gave informed written consent. The study was conducted in accordance with the Declaration of Helsinki and approved by the University of Trento Ethics Committee.

Procedure

In this experiment, participants were exposed to a stream of audio and visual stimuli while sitting in a magnetically shielded chamber. Auditory stimuli consisted of a compound of three pure tones (250 Hz, 500 Hz and 750 Hz), lasting for 50 ms and separated from each other by 50 ms, with three of the possible combinations (up: 250, 500, 750; down: 750, 500, 250; flat: 500, 500, 500). Visual stimuli were three different colored light (red, green and blue) presented via an LED positioned in front of the participants (Fig. 1A, 1.5° visual angle) with a duration of 250 ms. On each trial, auditory stimuli predicted the colored light according to the probability structure illustrated in Fig. 1B. The experiment was divided in three different sessions (Habituation, Acquisition and Extinction). What differed across the sessions was indeed the probability

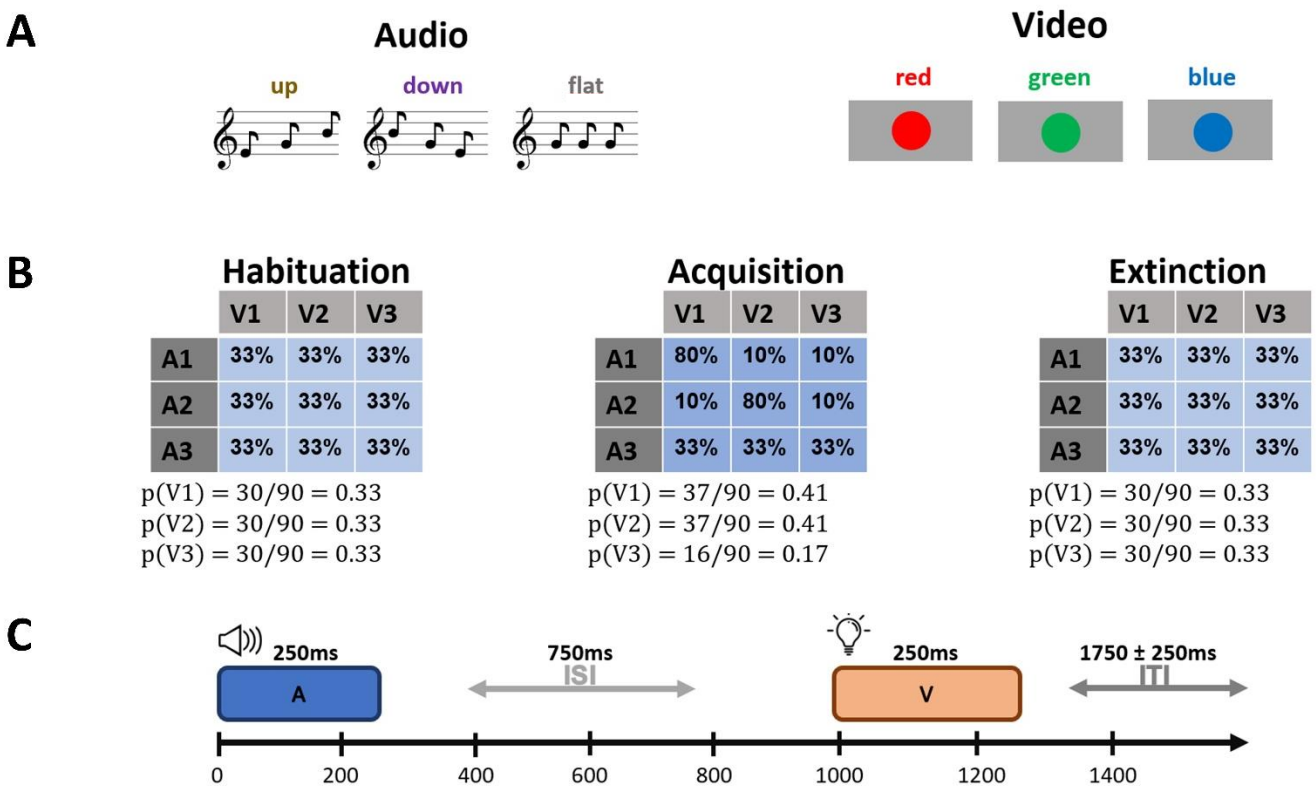

FIGURE 1 A. Stimuli presented during the experiment. **B.** Contingency table showing the percentage of occurrence of each visual outcome given an auditory stimulus for the three sessions Habituation, Acquisition and Extinction. **C.** Description of the trial structure.

structure of the audio-video stimuli association. In Habituation and Extinction, all audio-video associations had an equal probability of 33%. In Acquisition, one of the up or down tone compounds (A1) was paired with one of the three colors (V1) with a probability of 80% (V1|A1), while in the remaining 20% of the times, A1 was equally followed by the others colors. The other tone compound that was not selected as A1 was associated with another color different from V1 with the same probability structure as for V1,V2,V3|A1. In other words, V2|A2 was associated 80% of the trials and V1|A2 and V3|A2 only 10% each. The flat tone compound was associated with all colors with an equal probability of 33%. The assignment of the stimuli to the conditions was

counterbalanced across the participants. Only the A3 condition was always fixed to the flat tone compound, resulting in 12 possible combinations of audio-video associations. The trial structure, illustrated in Fig. 1C, consisted of the presentation of one of the three tone compounds, followed by an inter stimulus interval (ISI) of 1000 ms. After the ISI, one of the three colors was presented and then the trial terminated with an inter-trial interval (ITI) of 1750 ms ± 250 ms. The experiment consisted of 180 trials divided in 2 blocks for the Habituation and Extinction session each, and 360 trials divided in 4 blocks for the Acquisition. Critically, in order to investigate the role of task relevance in the modulation of cortical responses to sensory associations, we ask participants to perform a visual target detection task. The task consisted of pressing a button whenever they perceived the V1 color, thus making the V1|A1 association task-relevant while the V2|A2 task-irrelevant. Crucially, when debriefed at the end of the experiment with a questionnaire, participants did not become aware of both the relevant and irrelevant associations. The experimental script was generated using OpenSesame with PsychoPy as a backend (Mathôt et al., 2012).

MEG acquisition and preprocessing

We recorded MEG (Omega 2000, CTF Systems, Inc., Port Coquitlam, Canada) with 275 channels at a sampling rate of 1172.9 Hz in a magnetically shielded chamber. MEG data were preprocessed by firstly removing the DC offset subtracting the mean from each channel. Then we segment the data around the onset of the auditory compound selecting the time window from -300 ms to 2000 ms. After this step, we downsampled the data to 200 Hz and applied the robust polynomial detrending method to avoid the usage of the high-pass filter that can cause temporal artifacts. To apply the robust detrending, data were symmetrically mirror-padded with 100 seconds for each trial and then was first removed a linear trend followed by a 10^{th} order polynomial. Then, we applied

the Discrete Fourier Transform (DFT) filter in order to remove the line noise using as frequencies of interest 50 Hz and its harmonics up to 150 Hz. Artifact rejection was performed with a semi-automatic procedure by visual inspection and computing the variance of each trial and excluding which surpassed the threshold of 10^{-23}. Independent component analysis (ICA) was performed to get rid of the ocular and cardiac artifacts using the extended infomax algorithm. Finally, we baseline-corrected the data using the time window from -300 ms to 0 ms and low-pass filtered at 80 Hz using a Butterworth filter with order 4.

MEG analysis

In this study, we wanted to compare brain activity induced by relevant and irrelevant associations. Thus, we analyzed the time window going from the onset of the visual stimulus to 600 ms after this timepoint. To accomplish this goal, we employed Multivariate Pattern Analysis (MVPA) to decode differences in the V1|A1, V2|A2 and V3|A3 conditions between the Habituation and Extinction session. We applied linear discriminant analysis (LDA) with a 10-fold cross validation scheme and accuracy as performance metric to classify the trials belonging to the two sessions for each participant across time. We also ran searchlight analysis to understand which feature contributed most to the time-resolved performance. Furthermore, we also performed connectivity analysis to investigate functional networks underlying relevant and irrelevant conditions. For this reason, we first average the signal into different spatial regions of interest (ROI) according to the standard CTF montage, resulting in four ROIs (Frontal, Parietal, Temporal, Occipital). Then, we computed the time-frequency representation of the MEG data in each epoch of Habituation and Extinction of the three above selected conditions by convolving each ROI time-series with a set of complex Morlet wavelets and then taking the inverse Fast Fourier Transform. The wavelets were defined as $e^{i2\pi tf}e^{-t^2/(2\sigma^2)}$, where t is

time, *f* is frequency, and σ defines the width of each frequency band, set according to $n/(2\pi f)$, where *n* is the number of wavelet cycles. The frequency *f* increased from 8 to 45 Hz in 30 logarithmically spaced steps, and the number of cycles *n* increased from 3 to 12 in 30 logarithmically spaced steps. After computing the cross-spectral density, we computed the debiased weighted phase lag index (dWPLI; Vinck et al., 2011) across the time dimension, thus preserving the trial dimension. dWPLI is a measure of phase synchronization that is robust to the effects of volume conduction and uncorrelated noise and is not affected by the number of trials in each condition. This allowed us to run the searchlight analysis with the same hyperparameters as above but this time having as a spatial dimension the six ROI combinations. Finally, we averaged the results into 3 frequency bands (Alpha 8-13 Hz, Beta 13-25 Hz and Gamma 25-45 Hz). To assess the statistical significance of the classifier performance, cluster-based permutation tests were performed with an independent-samples t-test ($\alpha=0.05$), 10000 permutations and maxsum as cluster statistic.

Results

Statistical analyses of the LDA performance across time revealed a significant difference between V1|A1 and V3|A3 ($p < 0.05$, cluster corrected) and also between V2|A2 and V3|A3 ($p < 0.05$, cluster corrected) around the time window 340-390 ms. It should be noted that although all the three classifier performances are above the theoretical chance level, this cannot be interpreted as a significant result by itself. This is the rationale behind comparing the V1|A1 and V2|A2 performance against the V3|A3, because the former condition ensures a valid null model since there is no learning effect expected. Searchlight analysis conducted on the resulted significant time window revealed a major parieto-occipital contribution to the overall classification performance for the relevant and irrelevant condition, although for the relevant condition the spatial

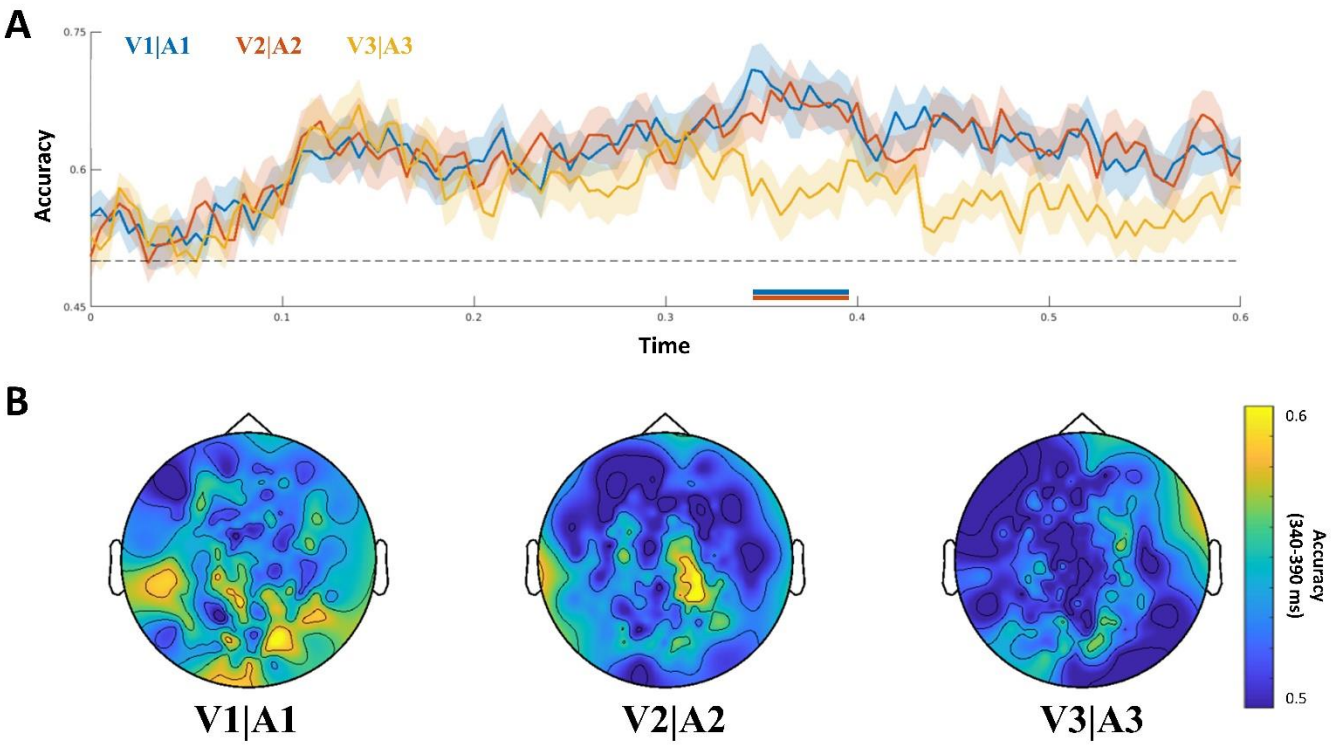

FIGURE 2 A. MVPA performance across time of V1|A1, V2|A2 and V3|A3 conditions. Shades indicate standard error. Horizontal bars indicate statistical significance with $p < 0.05$, cluster corrected. **B.** Topography of the searchlight analysis conducted on three conditions in the resulted significant time window 340-390 ms.

activation was more spreaded across the occipital region while for the irrelevant condition was more concentrated on the parietal region. After this result, we ran the functional connectivity analysis on the resulted significant time segment from the previous MVPA analysis across time and we found a significant difference between relevant vs neutral (V1|A1 vs V3|A3) and irrelevant vs neutral (V2|A2 vs V3|A3) in the gamma band across the Frontal-Parietal, Frontal-Occipital and Occipital-Parietal ROI combination (all $p < 0.05$, cluster corrected). These findings confirmed previous results on the time domain, as evidenced by the resulted occipital-parietal connectivity, and extended them showing that the frontal region played also a role in the encoding of the prediction error.

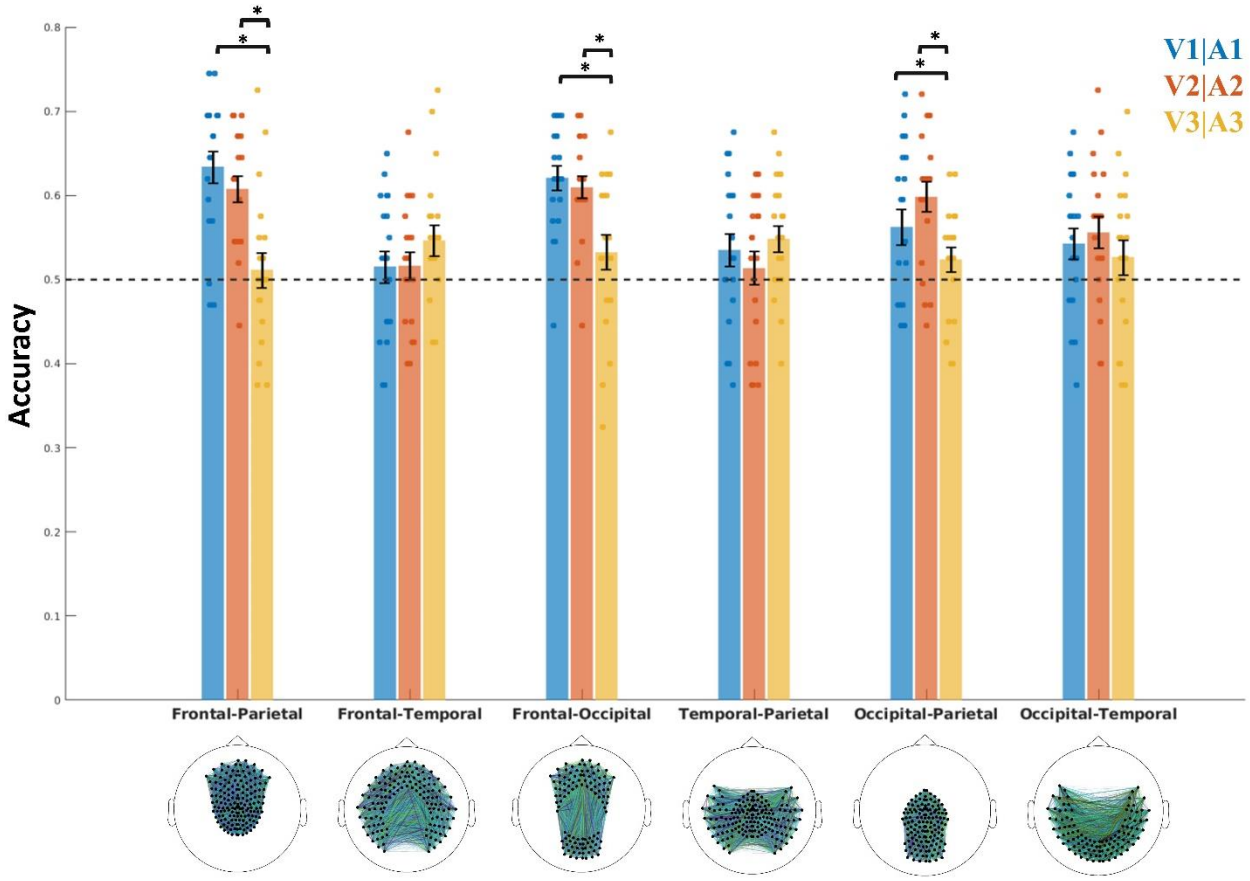

FIGURE 3. MVPA performance across combinations of ROI on the dWPLI connectivity data. Vertical bars indicate the performance of the LDA classifier of V1|A1, V2|A2 and V3|A3 conditions. Jittered points represented the classifier performance at the subject level. Error bars represent standard error. The horizontal dashed line indicates the theoretical chance level.

Discussion

In the present study, we investigated the time-locked activity and functional connectivity networks of the MEG signal underlying task-relevant and task-irrelevant associations. Participants were exposed to an audio-visual stream of stimuli while performing a perceptual detection task in which they had to press a button when perceiving the visual target, thus intentionally directing their attention to the cue-target association and making the other audio-visual associations irrelevant for the task they were instructed to perform. One of these

pairings (V2|A2) had the same probability structure of the cue-target association (V1|A1) across the experimental sessions, while the other (V3|A3) had a uniform probability structure across the entire experiment. One of the key findings of this study was that relevant and irrelevant associations had similar patterns of activation when comparing the brain responses to the onset of the visual stimulus. This can be interpreted as evidence that prediction errors are computed similarly regardless of the task relevance. Moreover, the latency of the effect found in this study for the task-irrelevant condition is similar to the effect found in the first study presented in this thesis with a shift of about 100 ms. This is also in line with previous robust findings in the literature about the neural signature of prediction error modulation such as P300 and Feedback Related Negativity waveforms. Regarding the comparison of these results with other studies in the literature that did not found a similar pattern for relevant and irrelevant associative learning (Stokes et al., 2014; Auksztulewicz et al., 2017), this can be explained by the application here of the multivariate analysis that is more sensitive to subtle pattern respect to the traditional methods implemented in previous studies. Indeed, for instance in Stokes et al. (2014) they just found an effect for the relevant condition but not for the irrelevant and the use of standard methods precluded them to robustly conclude that task relevance profoundly influences the way in which the brain computes sensory predictions and prediction errors. Another key result of this study was the functional coupling of fronto-parieto-occipital regions in the gamma band across the same time window of the results in the time domain. Critically, this effect was similar for both relevant and irrelevant associations. These findings on connectivity confirmed the previous results in the time domain from the searchlight analysis regarding the parieto-occipital coupling, and extend them showing the role of frontal regions in the processing of prediction error encoding. The involvement of the frontal lobe in the modulation of predictive processes has been

extensively reported in the literature (den Ouden et al., 2009; Dürschmid et al., 2016), therefore our results were largely expected. In addition, the pattern of oscilatory dynamics we found, the fronto-parieto-occipital coupling in the gamma band, can be interpreted under the recent evidence about the fact that different frequency bands may carry information about distinct stages of the predictive processing (Arnal & Giraud, 2012). The current framework about how predictive processes are encoded in the oscillatory dynamics of the brain is that there is an asymmetry between forward and backward passing of information (Bastos et al., 2015). In particular, ascending prediction errors are conveyed at higher frequency bands while descending predictions are encoded at lower frequencies (Friston, 2019). Our results are in line with this theoretical hypothesis since we found prediction error passing of information being carried on by gamma coupling of brain networks involving core areas for predictive processing such as frontal and parietal regions. Taken together, our results strongly suggested that, even if task relevance plays a role in the encoding of sensory prediction errors, the brain utilizes a general mechanism to predict incoming sensory data regardless of their imminent importance. A possible speculation about why this is the case could be the fact that the minimization of surprise is a meta goal for biological systems (Friston et al., 2006), therefore updating their internal models of the environment in order to predict potentially surprising events is also relevant (den Ouden et al., 2009). In other words, while task relevance can be regarded more as an extrinsic motivation that guides short-to-long term behavior, surprise minimization can be viewed as a strong intrinsic drive that motivates biological agents to act for long term behavior (Schwartenbeck et al., 2013). A limitation of this study was that we did not implement a computational modeling part in our pipeline analysis as well as we did not look at the estimated source space. Future directions may explore a more model-based approach to the comparison of relevant and irrelevant associations,

using both Bayesian or Reinforcement Learning models to account for prediction error patterns to fit the neuroimaging data, and implement a source localization method to be more precise about the spatial localization of the neural effects in the cortex.

How task relevance modulates brain activity

The aim of this thesis was to assess the role of task relevance in the modulation of brain dynamics during sensory associative learning, using a combination of computational modeling and multivariate pattern analysis in a range of associative learning tasks. In Chapter 1, I discussed previous literature suggesting sensory associative learning is mediated by the stimulus-stimulus contingency. Furthermore, prediction errors or surprising events, are thought to signal the need for updating the internal model of the environment, thus playing a central role for associative learning in biological systems. Indeed, surprise appears to be at the heart of not only to reward-based learning, but any form of associative learning. In Chapter 2, I presented a study in which participants were asked to perform a detection task while audio-visual stimuli were presented as distractors. These distractors were presented with a probability structure that made some of them more paired than others. Results showed that participants learned these task-irrelevant associations even without being aware of them. Moreover, occipital activity triggered by the conditioned auditory stimulus was elicited just before the arrival of the visual outcome and, after the onset of the unconditioned visual stimulus, a pattern of precision-weighted prediction errors estimated using an ideal Bayesian observer computational model correlated with EEG activity around 300 ms. In Chapter 3, we used the same task of the previous experiment adding two sessions before and after the main task in which all the conditional probabilities were identical for all stimuli pairs. Also, the experiment was spread across two weeks in six days. Results showed a difference in the modulation of the beta band induced by the presentation of the unconditioned visual stimulus preceded by the predictive and unpredictive conditioned auditory stimuli by comparing the pre and post sessions activity. In

Chapter 4, we investigated the time-locked activity and functional connectivity networks of the MEG signal by directly comparing task-relevant and task-irrelevant associations. Participants were exposed to an audio-visual stream of stimuli while performing a perceptual detection task, directing their attention to the cue-target association and making the other audio-visual associations irrelevant for the task they were instructed to perform. One of these pairings had the same probability structure of the cue-target association across the experimental sessions, while the other had a uniform probability structure across the all sessions. Results showed that relevant and irrelevant associations had similar patterns of activation when comparing the brain responses to the onset of the visual stimulus. Also, the activated functional networks were similar for both conditions with respect to the non associative condition and involved frontal, parietal and occipital regions. Taken toghether, these studies clearly demonstrate that, even if task relevance play a modulatory role on the strenght of the neural effects of associative learning, predictive processes take place in sensory associative learning regardless of task relevance. In particular, task-irrelevant associations resemble the same neural mechanisms found for relevant associations in both conditioned and uncoditioned related brain dynamics. Regarding the conditioned stimulus, we found evidence that a preparatory activity emerged for irrelevant predictions similarly as found in relevant contexts (Alink et al., 2010; Kok et al., 2012; Kok et al., 2017; St. John-Saaltink et al., 2015). This proactive, anticipatory activity is in line with predictive coding theories because it allows sensory cortices to be prepared for upcoming sensory data by efficiently preallocating neural resources (Kok et al., 2017). Regarding the unconditioned stimulus, we found a modulation of brain activity that followed a pattern of prediction errors computed from an ideal Bayesian observer, confirming that he same principles of learning by updating the internal generative model can be applied also to irrelevant associations. Interestingly,

the latency of the effect (300 ms) was also very similar to common patterns of event related components (ERP) found in the literature related to expectation paradigms (e.g. P-300 waveform). Moreover, by investigating the time-frequency representation of the unconditioned related neural activity, we found that beta band was more involved in processing these prediction errors and also this result is in accordance with common findings in the task-relevant literature. What is known from this literature is that there is an asymmetry between forward and backward passing of information in the oscillatory dynamics of predictive processes (Bastos et al., 2015). Specifically, ascending prediction errors are conveyed at a faster frequency, for example gamma band, while descending predictions are encoded at lower frequencies (Friston, 2019). The reason why we found a lower frequency band modulation due to prediction error encoding (beta) may be related to the fact that predictive processing are always active about multiple aspects of the sensory experience. Finally, by directly comparing relevant and irrelevant associations, we observed similar patterns of activations. Specifically, the latency of the decoded effect was very similar and around 350-400 ms, while the spatial topography was more different involving more frontal and occipital regions for the relevant and more parietal for the irrelevant. In other words, the spatial activation was more spread for the relevant condition with respect to the irrelevant, thus confirming that task relevance has a modulatory effect rather than completely changing the underlying mechanism behing associative learning. Furthermore, functional connectivity analysis confirmed the fronto-parieto-occipital involvement in the computation of prediction errors and resulted very similar for both conditions. Also, the effect was found in the gamma band for both conditions, additionally confirming the current view about oscillatory dynamics of predictive processing and extending to task-irrelevant contexts. A general theoretical interpretation about the findings that prediction errors are computed regardless task relevance can be

directly derived from the framework of predictive coding. In this framework of brain function, the critical goal for a biological system is the minimization of surprise (Friston et al., 2006). It is so fundamental, that can be conceived as a supra or meta goal for biological systems. Therefore, we can view the updating of the internal models of the environment (i.e. learning) the most relevant achievement because it allows to predict potentially surprising events in the future (den Ouden et al., 2009). In other words, we can see the fact that a biological system matches its predictions with sensory data, regardless of their imminent utility, as rewarding per se. Thus, while task relevance can be regarded more as an extrinsic motivation that guides short-to-long term behavior, surprise minimization can be viewed as a strong intrinsic drive that motivates biological agents to act for long term behavior (Schwartenbeck et al., 2013).

Limitations and future directions

In addition to the limitations of the specific designs and paradigms discussed in the results chapters, what follows are some general limitations about the methods considered in this thesis and future directions. One of the main limitation of the studies examined in this thesis is the poor spatial resolution given by the use of neuroimaging techniques such as EEG and MEG. This is due to the fact that in neuroimaging there is no method that can guarantee both temporal and spatial high resolution, therefore we opted for EEG and MEG because we wanted to better characterize the temporal aspects of task-irrelevant associative learning, since we capitalized our design on previous works done in fMRI and PET (den Ouden et al., 2009; McIntosh et al., 1998). Future studies on this direction can be considered using fMRI to better explore specific brain areas activity and interactions with other areas involved in sensory associations. Another limitation of these studies is the fact that we investigated only

functional connectivity (in the third study) without considering effective connectivity. The difference is that effective connectivity take into consideration also the direction of the information flow between two areas. There has been developed multiple methods to compute effective connectivity information such as Dynamic Causal Modeling (DCM, Kiebel et al., 2007) and Granger causality (Seth et al., 2015). The problem is that there are several pitfalls in the computation of effective connectivity in neuroimaging tools like EEG and MEG because of the volume conduction problem and the temporal and spatial correlation between the features. Future studies on this direction can take into account the implementation of effective connectivity measures that do not suffer from known problems (e.g. volume conduction). Finally, another limitation is that we use a bayesian computational model of associative learning to analyze EEG data (Hierarchical Gaussian Filter, in the first study). Although bayesian models are well known for being a good framework for modeling certain aspects of associative learning, the standard theoretical framework is reinforcement learning. Future studies can consider using reinforcement learning models such as the Rescorla Wagner model (Rescorla & Wagner, 1972) for trial-based estimation of the associative strength between conditioned and unconditioned stimuli or the Temporal Difference model (Sutton & Barto, 1981) for a real-time version of the estimated trajectories of the associative strength.

Bibliography

Akatsuka, K., Wasaka, T., Nakata, H., Kida, T., & Kakigi, R. (2007). The effect of stimulus probability on the somatosensory mismatch field. *Experimental Brain Research*, *181*(4), 607–614.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, *30*(8), 2960–2966.

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398.

Auksztulewicz, R., Friston, K. J., & Nobre, A. C. (2017). Task relevance modulates the behavioural and neural effects of sensory predictions. *PLOS Biology*, *15*(12), e2003143. https://doi.org/10.1371/journal.pbio.2003143

Baldeweg, T. (2006). Repetition effects to sounds: Evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences*.

Bastos, A. M., Litvak, V., Moran, R., Bosman, C. A., Fries, P., & Friston, K. J. (2015). A DCM study of spectral asymmetries in feedforward and feedback connections between visual areas V1 and V4 in the monkey. *Neuroimage*, *108*, 460–475.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bencsik, N., Pusztai, S., Borbély, S., Fekete, A., Dülk, M., Kis, V., Pesti, S., Vas, V., Sz\Hucs, A., & Buday, L. (2019). Dendritic spine morphology and memory formation depend on postsynaptic Caskin proteins. *Scientific Reports*, *9*(1), 1–16.

Bray, S., & O'Doherty, J. (2007). Neural Coding of Reward-Prediction Error Signals During Classical Conditioning With Attractive Faces. *Journal of Neurophysiology*, *97*(4), 3036–3045. https://doi.org/10.1152/jn.01211.2006

Brodt, S., Gais, S., Beck, J., Erb, M., Scheffler, K., & Schönauer, M. (2018). Fast track to the neocortex: A memory engram in the posterior parietal cortex. *Science*, *362*(6418), 1045–1048.

Cammann, R. (1990). Is there a mismatch negativity (MMN) in visual modality? *Behavioral and Brain Sciences*, *13*(2), 234–235.

Cichy, R. M., & Pantazis, D. (2017). Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. *NeuroImage*, *158*, 441–454.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Colas, J. T., Pauli, W. M., Larsen, T., Tyszka, J. M., & O'Doherty, J. P. (2017). Distinct prediction errors in mesostriatal circuits of the human brain mediate learning about the values of both states and actions: Evidence from high-resolution fMRI. *PLoS Computational Biology*, *13*(10), e1005810.

D’Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267.

Delamater, A. R., & Matthew Lattal, K. (2014). The study of associative learning: Mapping from psychological to neural levels of analysis. *Neurobiology of Learning and Memory*, *108*, 1–4. https://doi.org/10.1016/j.nlm.2013.12.006

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

den Ouden, H. E., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience*, *30*(9), 3210–3219.

den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A Dual Role for Prediction Error in Associative Learning. *Cerebral Cortex*, *19*(5), 1175–1185. https://doi.org/10.1093/cercor/bhn161

Domjan, M. (2005). Pavlovian conditioning: A functional perspective. *Annu. Rev. Psychol.*, *56*, 179–206.

Dürschmid, S., Edwards, E., Reichert, C., Dewar, C., Hinrichs, H., Heinze, H.-J., Kirsch, H. E., Dalal, S. S., Deouell, L. Y., & Knight, R. T. (2016). Hierarchy of prediction errors for auditory events in human temporal and frontal cortex. *Proceedings of the National Academy of Sciences*, *113*(24), 6755–6760.

Dürschmid, S., Reichert, C., Hinrichs, H., Heinze, H.-J., Kirsch, H. E., Knight, R. T., & Deouell, L. Y. (2019). Direct evidence for prediction signals in frontal cortex independent of prediction error. *Cerebral Cortex*, *29*(11), 4530–4538.

Eelen, P. (2018). Classical conditioning: Classical yet modern. *Psychologica Belgica*, *58*(1), 196.

Fasano, C., Bourque, M.-J., Lapointe, G., Leo, D., Thibault, D., Haber, M., Kortleven, C., DesGroseillers, L., Murai, K. K., & Trudeau, L.-É. (2013). Dopamine facilitates dendritic spine formation by cultured striatal medium spiny neurons through both D1 and D2 dopamine receptors. *Neuropharmacology*, *67*, 432–443.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., Papadakis, N., & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience*, *4*(10), 1043–1048.

Frankland, P. W., Josselyn, S. A., & Köhler, S. (2019). The neurobiological foundation of memory retrieval. *Nature Neuroscience*, *22*(10), 1576–1585. https://doi.org/10.1038/s41593-019-0493-1

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. J. (2019). Waves of prediction. *PLoS Biology*, *17*(10), e3000426.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Gallistel, C. R. (2002). Frequency, contingency and the information processing theory of conditioning. *Frequency Processing and Cognition*, 153–171.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*(3), 453–463.

Genoux, D., & Montgomery, J. M. (2007). Glutamate receptor plasticity at excitatory synapses in the brain. *Clinical and Experimental Pharmacology and Physiology*, *34*(10), 1058–1063.

Hauser, T. U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S., & Brem, S. (2014). Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry*, *71*(10), 1165–1173.

Hawkins, J., & Blakeslee, S. (2004). *On intelligence*. Macmillan.

He, J., Ding, L., Jiang, L., & Ma, L. (2014). Kernel ridge regression classification. *2014 International Joint Conference on Neural Networks (IJCNN)*, 2263–2267.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. J. Wiley; Chapman & Hall.

Helmholtz, H. (1925). *Treatise on physiological optics*. Rochester, NY: Optical Society of America.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *WIREs Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*(2), 519–530.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Ji, W., Suga, N., & Gao, E. (2005). Effects of agonists and antagonists of NMDA and ACh receptors on plasticity of bat auditory system elicited by fear conditioning. *Journal of Neurophysiology*, *94*(2), 1199–1211.

Jiao, H., Zhang, L., Gao, F., Lou, D., Zhang, J., & Xu, M. (2007). Dopamine D1 and D3 receptors oppositely regulate NMDA-and cocaine-induced MAPK signaling via NMDA receptor phosphorylation. *Journal of Neurochemistry*, *103*(2), 840–848.

Joel, D., & Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: An analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, *96*(3), 451–474.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In ba campbell & rm church (eds.), Punishment and aversive behavior (pp. 279-296). *New York: Appleton-Century-Crofts*.

Kiebel, S. J., Klöppel, S., Weiskopf, N., & Friston, K. J. (2007). Dynamic causal modeling: A generative model of slice timing in fMRI. *Neuroimage*, *34*(4), 1487–1496.

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270.

Kok, P., Mostert, P., & De Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, *114*(39), 10473–10478.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868.

Kumar, P., Goer, F., Murray, L., Dillon, D. G., Beltzer, M. L., Cohen, A. L., Brooks, N. H., & Pizzagalli, D. A. (2018). Impaired reward prediction error encoding

and striatal-midbrain connectivity in depression. *Neuropsychopharmacology*, *43*(7), 1581–1588.

Leuner, B., Falduto, J., & Shors, T. J. (2003). Associative memory formation increases the observation of dendritic spines in the hippocampus. *Journal of Neuroscience*, *23*(2), 659–665.

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*(1), 145–163.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, *275*(5297), 213–215.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825.

McIntosh, A. R., Cabeza, R. E., & Lobaugh, N. J. (1998). Analysis of Neural Interactions Explains the Activation of Occipital Cortex by an Auditory Stimulus. *Journal of Neurophysiology*, *80*(5), 2790–2796. https://doi.org/10.1152/jn.1998.80.5.2790

Mehta, M. R. (2001). Neuronal dynamics of predictive coding. *The Neuroscientist*, *7*(6), 490–495.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, *72*(2), 1024–1027.

Mirenowicz, J., & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, *379*(6564), 449–451.

Mullen, T. (2012). CleanLine EEGLAB plugin. *San Diego, CA: Neuroimaging Informatics Toolsand Resources Clearinghouse (NITRC)*.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251. https://doi.org/10.1007/BF00198477

Myers, N. E., Stokes, M. G., Walther, L., & Nobre, A. C. (2014). Oscillatory brain state predicts variability in working memory. *Journal of Neuroscience*, *34*(23), 7735–7743.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313–329.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660–E3669.

Pavlov. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Pazo-Alvarez, P., Cadaveira, F., & Amenedo, E. (2003). MMN in the visual modality: A review. *Biological Psychology*, *63*(3), 199–236.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*(1), 111–139.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197.

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rescorla, R. A. (1966). Predictability and number of pairings in Pavlovian fear conditioning. *Psychonomic Science*, *4*(11), 383–384.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current Research and Theory*, 64–99.

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, *63*(3), 592–606.

Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A., & Tonegawa, S. (2015). Engram cells retain memory under retrograde amnesia. *Science*, *348*(6238), 1007–1013. https://doi.org/10.1126/science.aaa5542

Salazar-Colocho, P., Del Rio, J., & Frechilla, D. (2007). Serotonin 5-HT 1A receptor activation prevents phosphorylation of NMDA receptor NR1 subunit in cerebral ischemia. *Journal of Physiology and Biochemistry*, *63*(3), 203–211.

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

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23.

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

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

Seth, A. K., Barrett, A. B., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, *35*(8), 3293–3297.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century.

Smith, A., Li, M., Becker, S., & Kapur, S. (2006). Dopamine, prediction error and associative learning: A model-based account. *Network: Computation in Neural Systems*, *17*(1), 61–84.

Squire, L. R. (2009). The Legacy of Patient H.M. for Neuroscience. *Neuron*, *61*(1), 6–9. https://doi.org/10.1016/j.neuron.2008.12.023

St. John-Saaltink, E., Utzerath, C., Kok, P., Lau, H. C., & De Lange, F. P. (2015). Expectation suppression in early visual cortex depends on task set. *PLoS One*, *10*(6), e0131172.

Stagg, C., Hindley, P., Tales, A., & Butler, S. (2004). Visual mismatch negativity: The detection of stimulus change. *Neuroreport*, *15*(4), 659–663.

Stefanics, G., Heinzle, J., Horváth, A. A., & Stephan, K. E. (2018). Visual mismatch and predictive coding: A computational single-trial ERP study. *Journal of Neuroscience*, *38*(16), 4020–4030.

Stokes, M. G., Myers, N. E., Turnbull, J., & Nobre, A. C. (2014). Preferential encoding of behaviorally relevant predictions revealed by EEG. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00687

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135–170. https://doi.org/10.1037/0033-295X.88.2.135

Tanaka, S., O'Doherty, J. P., & Sakagami, M. (2019). The cost of obtaining rewards enhances the reward prediction error signal of midbrain dopamine neurons. *Nature Communications*, *10*(1), 1–13.

Tazerart, S., Mitchell, D. E., Miranda-Rottmann, S., & Araya, R. (2020). A spike-timing-dependent plasticity rule for dendritic spines. *Nature Communications*, *11*(1), 1–16.

Terao, K., Matsumoto, Y., & Mizunami, M. (2015). Critical evidence for the prediction error theory in associative learning. *Scientific Reports*, *5*, 8929.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i.

Todorovic, A., Schoffelen, J.-M., van Ede, F., Maris, E., & de Lange, F. P. (2015). Temporal expectation and attention jointly modulate auditory oscillatory activity in the beta band. *PLoS One*, *10*(3), e0120288.

Treder, M. S. (2020). MVPA-Light: A classification and regression toolbox for multi-dimensional data. *Frontiers in Neuroscience*.

Treviño, M. (2016). Associative learning through acquired salience. *Frontiers in Behavioral Neuroscience*, *9*, 353.

Tye, K. M., Stuber, G. D., de Ridder, B., Bonci, A., & Janak, P. H. (2008). Rapid strengthening of thalamo-amygdala synapses mediates cue–reward learning. *Nature*, *453*(7199), 1253–1257.

van Ede, F., de Lange, F., Jensen, O., & Maris, E. (2011). Orienting attention to an upcoming tactile event involves a spatially and temporally specific modulation of sensorimotor alpha-and beta-band oscillations. *Journal of Neuroscience*, *31*(6), 2016–2024.

Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., & Pennartz, C. M. A. (2011). An improved index of phase-synchronization for electrophysiological

data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, *55*(4), 1548–1565. https://doi.org/10.1016/j.neuroimage.2011.01.055

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*(6842), 43–48.

Yerkes, R. M., & Morgulis, S. (1909). The method of Pawlow in animal psychology. *Psychological Bulletin*, *6*(8), 257.