frontiers in Computational Neuroscience

HYPOTHESIS AND THEORY
published: 14 June 2016
doi: 10.3389/fncom.2016.00054

A System Computational Model of Implicit Emotional Learning

Luca Puviani[1]* *and Sidita Rama*[2]

[1] *Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy,* [2] *Local Health Unit of Modena, Modena, Italy*

Nowadays, the experimental study of emotional learning is commonly based on classical conditioning paradigms and models, which have been thoroughly investigated in the last century. Unluckily, models based on classical conditioning are unable to explain or predict important psychophysiological phenomena, such as the failure of the extinction of emotional responses in certain circumstances (for instance, those observed in evaluative conditioning, in post-traumatic stress disorders and in panic attacks). In this manuscript, starting from the experimental results available from the literature, a computational model of implicit emotional learning based both on *prediction errors computation* and on *statistical inference* is developed. The model quantitatively predicts (a) the occurrence of evaluative conditioning, (b) the dynamics and the resistance-to-extinction of the traumatic emotional responses, (c) the mathematical relation between classical conditioning and unconditioned stimulus revaluation. Moreover, we discuss how the derived computational model can lead to the development of new animal models for resistant-to-extinction emotional reactions and novel methodologies of emotions modulation.

Keywords: amygdala, classical conditioning, emotional learning, evaluative conditioning, misattribution, prediction error, PTSD, UCS revaluation

OPEN ACCESS

Edited by:
Jose Manuel Ferrandez, Universidad Politecnica de Cartagena, Spain

Reviewed by:
Petia D. Koprinkova-Hristova, Bulgarian Academy of Sciences, Bulgaria
Dajiang Zhu, University of Southern California, USA

***Correspondence:**
Luca Puviani
luca.puviani@unimore.it

Received: *17 March 2016*
Accepted: *23 May 2016*
Published: *14 June 2016*

Citation:
Puviani L and Rama S (2016) A System Computational Model of Implicit Emotional Learning. Front. Comput. Neurosci. 10:54. doi: 10.3389/fncom.2016.00054

1. INTRODUCTION

In this manuscript, starting from a review and the analysis of the main experimental results in the field of implicit emotions, a novel interpretation of *associative learning* and *UCS revaluation* (and of their unavoidable interactions) is derived first. UCS revaluation represents the updating of the expected outcome (or biological value) associated with a given source of stimulation. In particular, if an UCS elicitation determines a greater (smaller) central nervous system (CNS) response with respect to the expected outcome, the value associated with the considered UCS will be increased (decreased) determining an inflation (deflation) process (Rescorla, 1974; Davey, 1989; Hosoba et al., 2001; Gottfried and Dolan, 2004; Schultz et al., 2013). Classical Conditioning occurs when an initial neutral stimulus (in other words a stimuli unable to activate the innate emotional system, so that it does not elicit emotional reactions, for instance a neutral sound) becomes paired to another stimuli, UCS, which elicits a biological relevant response, termed unconditioned response, UCR. After few CS-UCS pairings, the initial neutral CS becomes able to elicit a biologically relevant response, denoted conditioned response (CR) "similar" and generally speaking smaller than UCR (Fanselow and Poulos, 2005). In the literature CC and UCS revaluation are considered two independent learning mechanisms (Rescorla, 1974; Hosoba et al., 2001; Gottfried and Dolan, 2004). In this manuscript, considering that almost all the experiments reported in the technical

literature about implicit emotional learning involve discrete trials stimulation (e.g., electric shock delivery, food delivery) and measures (neuronal activity recordings, fMRI measures, or behavioral), the derived theory and model are initially defined in a discrete time scale. The proposed model is able to justify experimental results not predictable by other existing models, and it can be adopted for the study of important paradigms, such as the Iowa Gambling Task (Bechara et al., 1994; see Section 3.3). Furthermore, starting from the obtained discrete time model its continuous time counterpart is derived next. The derivation of such a continuous time model is based on mathematical considerations and engineering standard methods under the constraints imposed by the functional connectivity between the different brain regions involved in automatic emotional processing. A dynamical continuous time model which accounts for both (a) statistical/associative learning and pattern recognition and (b) for a time-varying stimulation intensity (i.e., implicit UCS revaluation) and the consistent related phenomena (e.g., the so called *emotional contrast effect*) has not been developed yet from our knowledge. This could be due to different reasons: first of all UCS revaluation has not obtained much attention over years and the researches have been focused mainly in CC; second, CC is intrinsically time-discrete. Nevertheless, the above cited continuous model is useful because shows the dynamics which lead to the updating of the emotional value over time, due for instance to a time-varying stimulation which exerts alternations of both aversive and appetitive values (for instance, a stimulation can elicit a slow aversive increase of tension and then a fast tension release, inducing a given organism to perceive it as an appetitive source of stimulation since it produces emotional rewarding effects). Indeed, classical conditioning model cannot describe the frequency or time emotional response under the influence of a time-varying stimulation, such an acoustic signal which varies between appetitive and aversive response induction (for instance varying both the sound frequency and intensity), as occurs in music. More specifically, a continuous time dynamical model can show how the emotional system *tracks* a given source of stimulation, either if such a source elicits the organism through an information flux (in other words the emotions are induced by aversive and appetitive information such as smiles or angry facial expressions, or a movie, but not by exerting a physical or energy based interactions) or through an energy based flux (i.e., through a stimulation due to energy exchange between the stimulus and the organism's receptors, such as a painful stimulation).

It is worth noting that emotional learning models which do not account for the implicit intensity stimulation evaluation (i.e., the UCS evaluation and revaluation over time or over trials) cannot predict or justify important psychophysiological phenomena which originates from specific dynamics of the emotional arousal. Such phenomena are the so called resistant-to-extinction (or inextinguishable) emotional responses, such as those observed in *evaluative conditioning* or in pathological reactions observed in panic attacks (Meuret et al., 2006) and post traumatic stress disorder (PTSD) (Beck and Sloan, 2012; Parsons and Ressler, 2013; Perusini et al., 2016). More specifically, emotional learning models based on associative learning (i.e., CC; Pavlov, 1927) account for the conditioned stimulus (CS) response variation due to the modulation of the statistical contingencies between an actively eliciting stimulus (called unconditioned stimulus, UCS) and the CS itself (which was neutral before the CS-UCS pairing), but they cannot say nothing about the intensity dynamics associated with the given UCS (which represents the causal source of stimulation). In other words, CC-based models describe the CS-UCS connection strength neglecting the relation between the UCS representation and the expected response associated with it (i.e., unconditioned response, UCR), which, in turn, may depend also (and indirectly) on the CS-UCS connection strength (see **Figure 1**). For these reasons, these models, cannot say nothing about (1) the neuronal populations involved in CS response (CR); (2) the mathematical expression of the intensity of the CR at the end of the acquisition process (in other words the CR intensity when CS predicts with absolute certainty the occurrence of UCS); it worth mentioning that until now the qualitative explanation is that the "*CR is similar but smaller than the UCR*" (Fanselow and Poulos, 2005); (3) the mathematical expression of UCR. The theory developed in this manuscript shed lights on these points, and, doing this, it will be able to justify how resistant-to-extinction emotional responses originate. Furthermore, the model permits the development of stimulation functions able to induce PTSD-like emotional reactions in animal models, or for emotional modulation (for instance decreasing an emotional response).

The manuscript is organized as follows. In the Paragraph "Materials and Methods" some fundamental definitions and concepts are provided and motivated first; thereafter, a review of the technical literature about empirical results and models on implicit emotional learning is presented, together with qualitative and quantitative considerations, which permit the development of the structure of the models (more specifically the description of the main brain regions involved and their functional role and connections) and, successively, the quantitative relations and constraints between the involved variables (more specifically the linearity hypothesis, the relation between a predicted/expected outcome and the reactive response, the *integration property* and the *emotional contrast effects*); successively, the main assumptions and hypothesis of the model are provided and motivated; in the subsequent Sections the model is developed in three different cases: (a) UCS revaluation in the discrete time scale; (b) joint CC and UCS revaluation in the discrete time scale; (c) continuous time scale. In the "Results" Section the main post-predicted results of the models and quantitative justification of specific psychophisiological phenomena are presented together with the comparison with existing models; furthermore, a Section on the validation, interpretation and applicability of the derived models and theory closes the Paragraph. Finally, the "Discussion" Section closes the manuscript.

2. MATERIALS AND METHODS

2.1. Definitions and Concepts

2.1.1. Motivation

This Section provides fundamental concepts and definitions for the development of the model (some of the following

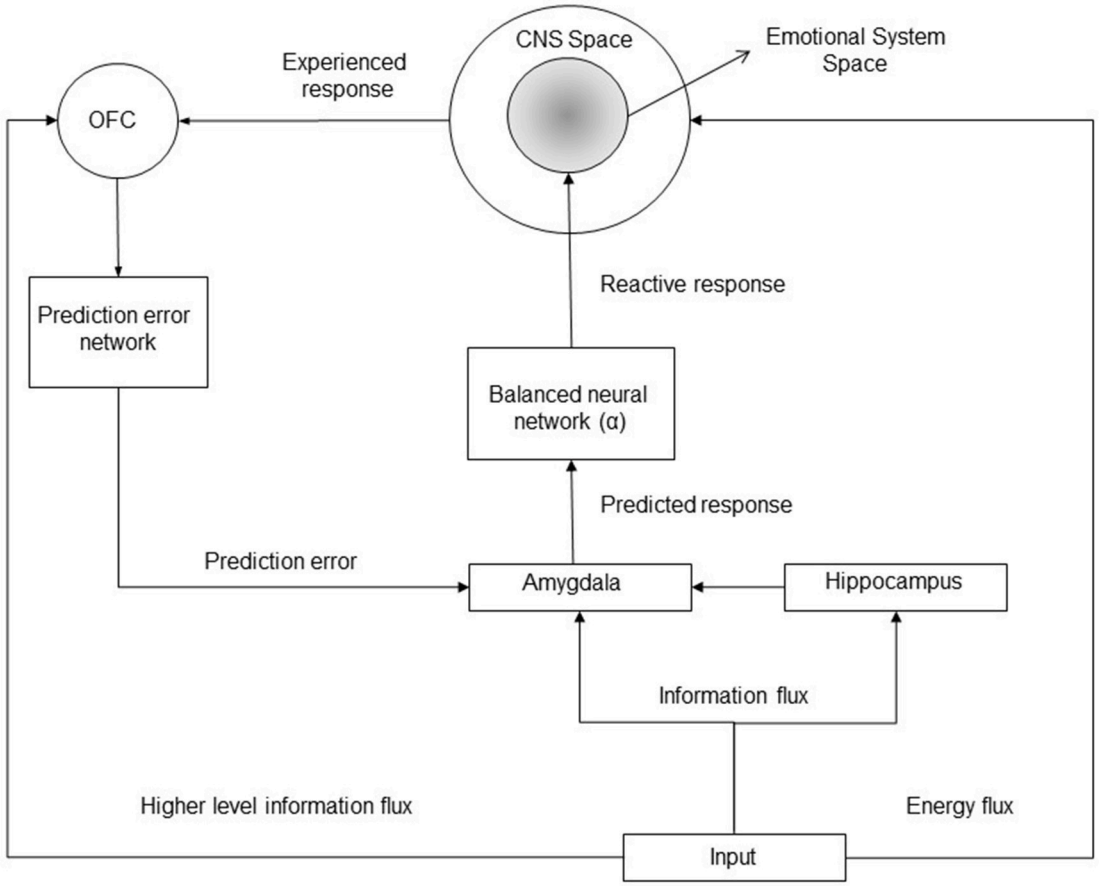

FIGURE 1 | Block diagram of the implicit emotional system. A given input to the emotional system can provide (1) an energy flux determining a cascade of reaction which finally elicit a CNS response (an increase/decrease of the mean firing rates of neuronal populations within the central nervous system); (2) an information flux determining a stimulus identification within the amygdala and within the orbitofrontal cortex (OFC); furthermore, specific higher cognitive level information can be processed by the OFC only. OFC encodes the biological predicted or expected outcome for a given stimulus, it manages the computation of prediction errors updating the expected emotional response (and the reactive response) within the amygdala. The amygdala elicits emotional neuronal populations with an intensity proportional to the expected response.

definitions and analysis are taken from Puviani et al., 2016). First a mathematical definition of CNS response is given, such a definition is useful for the subsequent definition of emotional (and reactive) response. These definitions are needed since, in CC theory and models, emotional responses (or reactions) are not well defined from a quantitative perspective; indeed in CC some behavioral or autonomic correlated responses are measured during experiments, such as the degree of salivation or the indirect measure of the arousal (i.e., the overall intensity of an emotional response) through the skin conductance response (SCR) evaluation. It is worth noting that such indirect measures always reflect a CNS response. Furthermore, the definition of *source of stimulation* is provided, since in CC theory the distinction between CS and UCS is based on the fact that an UCS exerts an innate reaction and the CS does not; nevertheless, this differentiation is not always satisfactory, since a previously neutral stimulus could became an UCS in certain circumstances. Moreover, the differentiation of two different types of stimulation are specified, these are the *active* and *reactive* stimulations; conversely in CC theory and derived models this distinction is not considered, so that it is not possible to express the overall CNS response as the contribution of the two quantities which, generally speaking, can vary independently during emotional learning (indeed, the brain integrates the two contributions, so that they becomes indistinguishable from a neurophysiological perspective, but the analytical distinction is useful from a model perspective). Finally, the definition of *reactive mimicking* property is provided and it is based on empirical evidences from pharmacological conditioning experiments. Such a property shed lights on the emotional (and, generally speaking, the reactive) learning mechanism, evidencing that whenever an UCS stimulates a CNS response, the brain stores the reactive response (i.e., the intensity and the elicited CNS sub-population) which will be associated with the given UCS for future predictions.

2.1.2. Definitions

2.1.2.1. CNS response

A generic response induced within the CNS can be represented by the superposition of the activity of different neural populations;

more specifically, assuming that the CNS consists of N different populations, the response, denoted with the vector $\mathbf{y}$, can be expressed as:

$$\mathbf{y} = \sum_{i=1}^{N} y_i \mathbf{v}_i, \quad (1)$$

where y_i represents the i-th neuronal population activity and $\{\mathbf{v}_i; i = 1, 2, ..., N\}$ represents a set of versors, being associated with different neuronal populations, which form a complete basis $\mathcal{B}$ for the CNS space. More specifically, y_i is a real quantity representing the product between the mean number of elicited neurons and their mean firing rates for the i-th neuronal population (with $i = 1, 2, ..., N$); consequently, y_i takes on a positive (negative) value if the response produces an increase of (a decrease or inhibition of) the activity for the i-th population, and is equal to zero whenever the response does not involve any adjustment for the baseline activity of the population. It is worth noting that the different neuronal populations could be interdependent (i.e., $\mathcal{B}$ does not represent an orthonormal basis).

2.1.2.2. Source of stimulation

A "source of stimulation" is defined as any stimulus able to causally and directly induce a CNS response (e.g., a painful stimulation). Some sources of stimulation (more specifically their neural representations) are natively coded within the mammalian brain, shaped by evolution (Ohman, 1993; Ohman and Soares, 1993; Esteves et al., 1994), while others are acquired through experience (Flykt et al., 2007); nevertheless a conditioned stimulus does not represent a source of stimulation, since it cannot causally and directly determine a CNS response, instead it may signal an imminent stimulation of a given UCS, and, for this reason, it can indirectly determine a CNS response. It can be inferred from experimental results based on subliminal stimulation (Ohman and Soares, 1993) that encoding a stimulus as a source of stimulation (i.e., as the responsible of the elicitation) or as contextual or conditioned stimulus, makes the difference in the determination of the specific brain region in which it will be stored; more specifically only the sources of stimulation are stored in the *basolateral amygdala* (BLA) in a rapid-access region, elicitable through the thalamo-amygdala pathway, while CSs do not. The terms "source of stimulation" and "UCS" are adopted indiscriminately in the following.

2.1.2.3. Active stimulation and active response

An active stimulation is defined as any stimulation causally and directly exerted by an UCS *through an energy flux* (e.g., mechanical, thermal, chemical, pharmacological...) exchanged between the UCS itself and a given organism. Whenever an UCS exerts an active stimulation the resulting elicited CNS response is causally and directly related to the intensity of the energy flux (and its temporal derivatives) transferred from the source of stimulation to the organism's receptors. For instance a painful thermal stimulus exerts an active painful elicitation transferring heat to specific receptors of the given organism; furthermore if the transferred heat increases the perceived painful response will increase too.

2.1.2.4. Reactive stimulation and reactive response

A reactive stimulation is defined as any stimulation induced within the CNS exclusively *through an information flux*. Thus, a reactive stimulation exerts its action through information processing and not by a direct energy flux. For instance, a CS previously paired with a given UCS induces a CNS response (e.g., fear) through its mere perception (i.e., information processing) and not because energy flux transfer toward the organism. It is important pointing out that, obviously, the mere perception of a stimulus (e.g., a CS) is sustained by a certain energy flux, such as acoustic (mechanical) or light intensity variation (electromagnetic), nonetheless, in this case, the response induced within the CNS is not causally and directly determined by the energy flux, or, in other words, the response is not directly related to the intensity of the energy flux, instead here the energy represents a mean to transfer an information flux. Indeed, a CS, may determine a CNS response because the information flux revealing its presence triggers a previously learned response. Generally speaking, a reactive response consists of a "self-induced" reaction triggered by an information flux (e.g., by a visual, auditory, olfactory, gustatory perception or by imagination), conversely, an active response is sustained by an external energy flux (e.g., an active drug or an electric shock). It is possible to exert both active and reactive stimulations concurrently; for instance, an hidden drug administration exerts only an active stimulation, since a pharmacological (chemical) flux is provided while no information is given, conversely, an open drug administration may induce both an active pharmacological response and a reactive stimulation due to cognitive (and even unconscious or imaginary) information processing (Amanzio and Benedetti, 1999; Benedetti et al., 2003; Benedetti, 2008).

On the basis of the above mentioned definitions it follows that an UCS can exert both an active and a reactive stimulation, while a CS can induce only a reactive stimulation.

2.1.2.5. Reactive (emotional) system

Generally speaking, a reactive stimulation cannot involve all the CNS neural components, since, for instance, a somatosensory stimulation can occur only through an energy flux (e.g., mechanical) and not by a simple information processing; for this reason only a "sub-space" of the CNS neuronal populations can be reactively elicited. The CNS sub-space which can be elicited through a reactive (information flux based) stimulation is termed reactive system (as will be clarified in the following the emotional system represents a sub-space of the reactive system). Hence, provided that N denotes the number of the distinct neuronal populations within the CNS (Equation 1), and that K denotes the number of the reactive system neural populations, it follows that $K \in N$. Which are the neuronal components within the CNS belonging to the reactive system? The answer comes from classical conditioning and pharmacological conditioning experiments in which a CS exerts a reactive stimulation after being paired with an active UCS. From the technical literature emerges that the reactive system may involve: (1) emotional responses (which include, for instance, the dopaminergic mesolimbic and mesocortical system, Scott

et al., 2007; Colloca, 2014; the fear and anxiety related circuits, McNally et al., 2011; Li and McNally, 2014; the endocannabinoid and opioid system in placebo analgesia, De Pascalis et al., 2002; Petrovic et al., 2002; Zubieta et al., 2005; Wager et al., 2007; Eippert et al., 2009; Watson et al., 2009; Nolan et al., 2012, the serotoninergic system, the target neuronal systems of depression, anxiety and addiction; see Benedetti, 2008); (2) the dopaminergic motor system (De la Fuente-Fernandez et al., 2001; De la Fuente-Fernandez and Stoessl, 2002); (3) the *humoral immune response system* (in particular the components of the CNS such as the hypothalamic-pituitary-adrenal axis, HPA, or the sympathetic nervous system, SNS; Goebel et al., 2002; Cacioppo et al., 2007; Benedetti, 2008; Vits et al., 2011); (4) the endocrine system; (see Benedetti, 2008; Enck et al., 2008 for a review).

2.1.2.6. Reactive (and emotional) mimicking

From a growing body of literature (Amanzio and Benedetti, 1999; Petrovic et al., 2002; Haour, 2005; Eippert et al., 2009; Guo et al., 2010; Lui et al., 2010; Nolan et al., 2012) it is reported that pharmacological conditioning determines a reactive response which mimic the active pharmacological response. The above mentioned property is termed here reactive mimicking. For instance, experimental results reported in Ito et al. (2000) show that an increase in dopamine release in the *ventral striatum*, measured through microdialysis, are observed not only when rats self administer cocaine (UCS), but also when they are solely presented with a tone (CS) that has been previously paired with cocaine administration. Furthermore, provided that the reactive system represents only a subset of the CNS, it is evident that only such a subset of the CNS neuronal populations can be mimicked. For instance, a CS previously paired with a painful UCS stimulation will be able to elicit only a specific portion of the components that were actively stimulated by the UCS; such components represent the emotional response (e.g., the activation of anterior cingulate cortex and the anterior insula; Singer, 2004), and, they cannot involve the somatosensory neural populations, even if these were involved in the original UCR.

2.2. Derivation of the Emotional Dynamical System Structure

In this Section the role of the key brain regions involved in emotional processing and response are reviewed from the literature. The purpose of this Section is to infer the functional structure of the dynamical emotional system.

2.2.1. Emotional Responses, Amygdala and Orbitofrontal Cortex

In mammalian brains the amygdala represents the core center in the formation and storage of emotional events and in the elicitation of emotional responses. In particular, in a growing body of literature (Schoenbaum et al., 1999; Glascher and Adolphs, 2003; Paton et al., 2006; Choi and Jeansok, 2010; Amano et al., 2011; Sangha et al., 2013) it is shown that amygdala is necessary for fear responses, and that no reactive fear responses are instantiated in the absence of an intact amygdala (Choi and Jeansok, 2010). Furthermore, the amygdala mediates both appetitive (i.e., rewarding) and aversive stimuli (Muramoto et al., 1993; Schoenbaum et al., 1999; Paton et al., 2006; Shabel and Janak, 2009; Amano et al., 2011; Sangha et al., 2013; Gore et al., 2015); in the former case the *basolateral amygdala* (BLA) neurons project onto the nucleus accumbens (NAcc), whereas in the latter one onto the *centromedial amygdala* (CeM) (Namburi et al., 2015). Hence the amygdala represents the key region for the elicitation of any reactive emotional response and it elicits (both directly or indirectly) emotional and motivational areas of the brain (LeDoux, 2000; Sah et al., 2003; Gore et al., 2015; Janak and Tye, 2015; Tovote et al., 2015). Nevertheless, it is worth noting that if the amygdala is damaged, an active elicited response (e.g., an unconditioned painful stimulus) can be still elicited. Moreover, experiments performed adopting optogenetic manipulations have evidenced that the representation of any UCS is stored within the BLA (Redondo et al., 2014; Gore et al., 2015). However, further fMRI studies (Gottfried et al., 2003; Gottfried and Dolan, 2004; O'Doherty, 2004; Kringelbach, 2005; Dolan, 2007; Pessoa, 2010) have shown that UCS representations (and its associated "biological values" or, in other words, the outcome which is expected from the given UCS) are encoded not only within the amygdala, but also in the orbitofrontal cortex (OFC). The fact that a stimulus representation and its associated expected outcome are stored in different brain regions (i.e., duplicated) could seem a waste of resources; nevertheless different reasons could justify this redundancy. Indeed, on the one hand it is important that the representation of relevant stimuli (such as fear relevant stimuli) are accessible through rapid access pathways, such as thalamo-amygdala pathway (LeDoux, 1996, 2000; Ohman, 2005), promoting a quick reaction whenever the stimulus is perceived. On the other hand, it is also important that a stimuli representation can be integrated with relevant cognitive information (when available) for the inference or prediction of the probable outcome. For instance, animals may learn that a given stimulus (e.g., a predator) is threatening observing others facing with it (Olsson et al., 2007; Olsson and Phelps, 2007), without the need of experiencing directly a stimulus elicitation. Hence, the OFC integrates different pieces of information (especially higher level cognitive ones) for inferring a probable outcome, and to update the response associated with the given UCS in "faster" subcortical regions (i.e., in the amygdala). Furthermore, prefrontal regions, like the dorsolateral prefrontal cortex (DLPFC), may interact with OFC to enhance or inhibit the response elicited by the amygdala (Ohman, 2005; Dolan, 2007). For instance, initial amygdala response to a fear-relevant but non-feared stimulus (e.g., pictures of spiders for a snake phobic) disappears with conscious processing by the activation of DLPFC and OFC (Ohman, 2005). Furthermore, also experiments in the filed of decision making have evidenced that OFC supervises the amygdala (Wallis, 2007; Rolls and Grabenhorst, 2008; Kennerley and Walton, 2011). Finally, it is worth pointing out that OFC is not necessary for classical conditioning, however, it is certainly needed for modifying the response if the predicted outcome is revaluated (i.e., UCS inflation and devaluation; Gallagher et al., 1999; Stalnaker et al., 2015).

2.2.2. Error-Driven Learning

From a growing body of literature emerges that learning occurs through the computation of specific *error-signals* (or *prediction errors*) (Schultz and Dickinson, 2000; Garrison et al., 2013). Generally speaking, the prediction error is defined as the difference between the response (or the outcome) expected from a given stimulation and the response actually perceived by the elicited organism. This definition relies on experimental observations acquired in functional imaging studies (Berns et al., 2001; O'Doherty et al., 2003; Garrison et al., 2013), or directly measured in dopaminergic circuits (e.g., in the *ventral tegmental area*, VTA) or in other fear-related circuits (Schultz, 2000, 2006; Schultz and Dickinson, 2000; Waelti et al., 2001; Bray and O'Doherty, 2007; Delgado et al., 2008; McNally et al., 2011; Steinberg et al., 2013; Li and McNally, 2014).

Different mathematical models describing classical conditioning learning (e.g., Rescorla-Wagner model, Rescorla and Wagener, 1972; Miller et al., 1995, or *temporal difference* (TD) *models,* Sutton, 1988; Sutton and Barto, 1990; Schultz et al., 1997; O'Doherty et al., 2003), or describing learning in general, such as the probabilistic (Bayesian) "perception" and "action" learning models (i.e., the *predictive coding* (PC) (Friston, 2003, 2008) and *active inference model* (Friston et al., 2009, 2010), assume that coding behavioral responses involves the computation of a prediction error. More specifically, the brain makes predictions in relation to a given stimulus and, on the basis of the experienced outcome, the prediction is updated through the prediction error. If the experienced outcome is greater (lower) than the prediction, the computed error signal is positive (negative) and corrects the new prediction; furthermore, if the experienced response coincides with the expected outcome, the error signal is zero and no prediction updatings take place.

2.2.3. On the Computation of the Prediction Errors

A growing body of literature (Schultz, 1998, 2000; Waelti et al., 2001; Schultz, 2006; Delgado et al., 2008; Bourdy and Barrot, 2012) evidenced that in emotional learning, populations of dopaminergic neurons encode prediction errors evaluating the difference between what is expected (i.e., the expected reward) and what is really occurring; furthermore, the prediction error is exploited to correct and modulate the individual's emotional and behavioral response. The prediction error computed in these dopaminergic regions can be positive or negative and can drive appetitive or aversive emotional reactions (Delgado et al., 2008).

It is not completely clear if prediction errors driving emotional responses are evaluated in different brain regions, depending on the nature of the involved emotional neuronal populations, or if dopamine neurons encode prediction errors related to all the involved populations; however, in the computation of the emotional error signal, a fundamental role is played by the OFC (O'Doherty, 2007). In fact, various experimental results have evidenced that the OFC generates information about expected outcomes which are deemed critical in the computation of prediction errors (e.g., see Takahashi et al., 2009 and references therein) and these results are consistent with the relation between the reward-related activity in OFC and VTA dopamine neurons (Takahashi et al., 2009). Experimental results have also evidenced that, when OFC and midbrain data are juxtaposed, anticipatory activity observed in the OFC is inversely related to dopaminergic error signaling downstream (Stalnaker et al., 2015). This suggests that the error signals in other brain areas might depend partly on OFC input for properly calculating the errors (Schoenbaum et al., 2009; Stalnaker et al., 2015).

2.2.4. The Role of the Hippocampus in Emotional Learning and Biological and Functional Differences between UCS Revaluation and Classical Conditioning

As reviewed above, the amygdala encodes the representation of UCSs and the related emotional responses; furthermore, it is well known that contextual information and statistical contingencies associated with a given UCS are encoded by the hippocampus (Bechara et al., 1995; Richardson et al., 2004). Important questions arise: what is the functional connectivity between a CS stored in the hippocampus and an emotional response? Does the hippocampus store emotional responses associated with the CSs? Which is the functional connection between the hippocampus and the amygdala? Responding to these questions permits to elucidate the role of the hippocampus in emotional learning and to differentiate the two learning mechanisms: CC and UCS revaluation. Such responses come from recent optogenetic experimental results (Redondo et al., 2014; Gore et al., 2015) which have evidenced that the hippocampal engram memory (which codes a CS) is neutral and could freely associate with either positive or negative emotions, through the UCS representation coded within the BLA. Furthermore, optogenetic reactivation of the hippocampal *dentate gyrus* (DG) engram cells coding a CS, during the presentation of a new UCS having valence opposite to the original UCS (which was previously paired with the CS itself), strengthens the connectivity of these cells with the new subset of the BLA neurons, while weakening the connections established during the original learning process. In other words, the simultaneous activation of a CS neural representation and of a new UCS strengthens a CS-UCS synaptic connection and, at the same time, weakens the connection between the CS and the previously associated UCS, which is not simultaneously active. These results evidence three important features: (1) a CS stored in the hippocampus has to be connected to an UCS representation within the BLA in order to trigger an emotional response; (2) the CS-UCS connection can be strengthened or weakened through synaptic Hebbian plasticity (i.e., through the mechanism "cells that fire together wire together," without the need of error signal computations or UCS revaluation); (3) the fact that the CS engram memory is emotionally neutral it means that the emotional reaction triggered whenever it is perceived is exclusively due to the CS-UCS *synaptic strength,* denoted ω_{CS-UCS} in the following. In turn, the UCS representation is associated with an emotional value (denoted i_R in the following). Hence, the term i_R represents the *reactive response* triggered whenever a CS connected with the given UCS is perceived.

Other experimental evidences support the fact that CC is not driven by prediction errors (see Section 3.2.4).

2.2.5. Functional Connectivity of the Implicit Emotional System

On the basis of the reviewed results in the previous Sections (more specifically see Sections 2.2.1–2.2.4) the functional connectivity of the brain regions involved in the implicit emotional learning can be inferred (see **Figure 1**). In particular, it is shown that a given stimulation can elicitate both an information and an energy flux, more specifically, the energy flux determines a direct response within the CNS system (for instance a painful stimulation determines an increasing firing rates of the neurons belonging to the insula, the anterior cingulate cortex, the sensorimotor cortex, and others), while the information flux can be processed by the amygdala (e.g., by a mere stimulus perception), by the hippocampus (for statistical and contextual recognition) and by the OFC (which can process higher level and structured information). The amygdala elicits a reactive response onto the emotional system (which involve only a sub-population of the entire CNS neuronal populations), and such a response can be modulated and corrected through the error signals whose computation is managed by the OFC.

Moreover, the reviewed results in Section 2.2.4 permits to infer the representation of CC and UCS revaluation as sketched in **Figure 2**. In particular, it is shown that UCS is stored within the amygdala, which in turn projects (through direct and indirect systems as reviewed in Section 2.2.1) onto the emotional system within the CNS; furthermore, the amygdala response is modulated by prediction errors with a feedback loop, which determines the UCS revaluation. On the other hand the CS (which, depending on the type, can be stored within the hippocampus or even in a region of the amygdala different from the region which contains the UCS representations) is not directly associated to an emotional response, but it is connected with an UCS representation, whose connection strength is denoted ω_{CS-UCS}. Moreover, it can be shown that UCS revaluation does not change the CS-UCS synaptic connection strength, conversely, as it is clarified in (Section 2.5.2; see also "Supplementary Material") an increase of CS-UCS connection strength leads inevitably to an UCS inflation until i_R reaches an asymptotic value.

2.2.6. Attribution of a Source of Emotional Stimulation and Predictive Coding

In the previous Sections it has been shown that a given UCS representation within the BLA is being associated with a specific reactive response which can also be modulated through prediction errors (see **Figure 2**). In order to build such a structure, the brain has to infer which is the stimulus that generates the CNS response. For instance, a device eliciting a painful response will be encoded as an UCS since it can be easily detected and attributed as the causal *source* of the painful stimulation; in this case the source of stimulation has been attributed correctly. However, whenever an emotional response due to a source of stimulation is attributed to a "wrong" source, an event of *source misattribution* occurs (Cotton, 1981; Bryant, 2003; Jones et al., 2009 and references therein). It is worth pointing out that misattribution may result either from conscious, accessible and measurable controlled processes, or from spontaneous, inaccessible, automatic processes (Uleman,

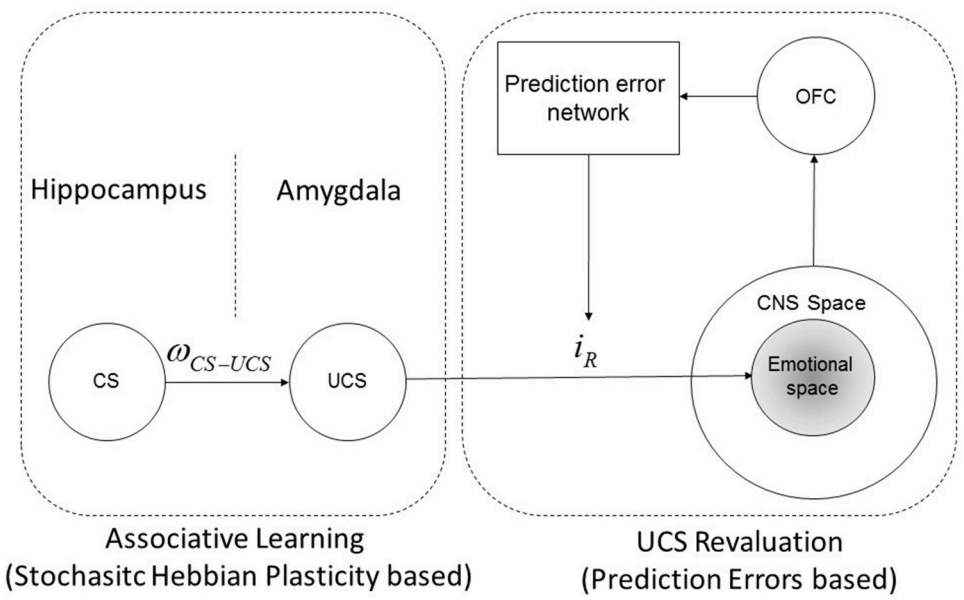

FIGURE 2 | CS-UCS-emotional response diagram. Schematic representation of the synaptic connections between a CS, the associated UCS and the reactive response (i_R) associated with the representation of the UCS itself. The reactive response is determined and modulated by UCS revaluation learning, instead the connection strength between CS and UCS (ω_{CS-UCS}) is determined by classical conditioning learning. The reactive response acts within a central nervous system sub-space called emotional space, which involve the emotional neuronal populations. Note that this schematic representation is in agreement with the experimental results shown in Gore et al. (2015) and Redondo et al. (2014).

1987; Anderson, 1989). In the last case this phenomenon is called *implicit misattribution* (Uleman, 1987; Anderson, 1989; Hutter and Sweldens, 2013). Generally speaking, the brain is an inference machine that actively predicts and explains its sensations; more specifically, the brain tries to explain the cause of its sensations through a probabilistic model (Friston, 2010). This concept is at the basis of the Bayesian brain hypothesis and of the so called *predictive coding* theory (Friston, 2008), which shows how automatic inference about the causes of sensory inputs are performed in a hierarchical structure within the brain. What is important in the development of our model is that *UCS attribution*is based on complex hierarchical and recursive (feedforward and backward) signals propagation between different layers, which generate a *probabilistic* model and representation about the cause(s) of the stimulation. This means that in the structure of the model we derived (see **Figure 2**), the UCS-i_R association is subject to eventual misattributions, and that two or more UCS representations could be associated to a (shared) given response if the brain fails to correctly infer the actual eliciting UCS; furthermore, in a limit case, a neutral and irrelevant stimulus could be attributed as the source of stimulation (i.e., a misattribution occurs). The attribution and misattribution phenomena can be quantitatively considered in the general structure of our model. In fact, it can be assumed that the reactive response the brain associates to a given UCS is proportional to the degree of *cause attribution belief* the brain predicts for that UCS. For instance, in the presence of two possible eliciting stimuli, the brain can infer all the possible probability attributions between the ranges (0–100%) and (100–0%), and, such a cause probability assignment is determined by (a) Bayesian prior belief distributions and (b) the actual stimulation conditions and perceptions. The attribution and misattribution phenomena will play an important role for the quantitative explanation of *evaluative conditioning* (see Section 3.2.3).

2.3. Quantitative Relations and Constraints between Variables

In this Section the main quantitative/mathematical assumptions and the relations between the variables involved in implicit emotional learning are inferred by a review of the literature and through analytical considerations.

2.3.1. Linearity Hypothesis

Recent computational and *in vivo* analysis have evidenced that cortical circuit have recurrent excitatory and inhibitory connections (van Vreeswijk and Sompolinsky, 1996, 1998; Doiron and Litwin-Kumar, 2014; Pehlevan and Sompolinsky, 2014; Deneve and Machens, 2016). Such a network architecture comprises excitatory and inhibitory neuronal populations, and the connectivity could be random and sparse. Computational studies about large networks reveal that the dynamics tends to a natural stationary state called *balanced state*. In this state, a balance between the excitatory and inhibitory inputs emerges dynamically for a wide range of parameters, and the internal synaptic inputs act as a strong negative feedback, which linearizes the population responses to the external drive despite the strong non-linearity of the individual cells. This feedback also greatly stabilizes the system's state and enables it to track a time-dependent input on time scales much shorter than the time constant of a single cell (van Vreeswijk and Sompolinsky, 1998). Hence a balanced network configuration not only stabilizes, linearizes (and makes deterministic) the input-output transfer function, but also makes the network capable of fast tracking of temporal changes in the input.

It is worth noting that in a balanced network configuration a linear behavior emerges from the chaotic behavior of individual neurons, so that chaotic balanced networks can precisely track any input signals, and the tracked signals can be read out by averaging spikes over the whole network population (van Vreeswijk and Sompolinsky, 1996). Therefore, at a system level (i.e., considering large brain region networks) if a CNS response elicitation is considered as an input, the further network processing can be considered linear. Nonetheless the relation between an external energy flux and the corresponding CNS neuronal response is generally non-linear (e.g., an acoustic stimulation). This means that linear modeling techniques and methods can be adopted provided that the considered system input is represented by a given CNS neuronal population activity (see Equation 1), and, the (non-linear) mapping between an external stimulation and the corresponding CNS response has to be further derived when needed, understood as a separate issue.

2.3.2. Neurophysiological *Integration Property* of Active and Reactive Contributions

In this subsection we review empirical evidence about the property of the brain of integration of active and reactive contributions. In practice we argue that the overall response induced within the CNS is always determined by an active (energy based) and a reactive (pure information) contribution, and that the brain cannot discriminate between these two quantities during the response (or outcome) evaluation. This fact could also represent the basis of the placebo induced response (Puviani and Rama 2016).

Experimental verification of the influence of nonconscious conditioned stimuli on placebo/nocebo effects (Jensen et al., 2012, 2015) show that a reactive stimulus is able to interfere with a given active stimulation (e.g., an active drug or a painful stimulation), by increasing or decreasing the effect of the active response. This suggests that common active and reactive response components can be additive or competing and, hence, both contribute to the determination of the overall elicited response within the CNS. This observation is supported by further experimental results (Roy et al., 2009; Wagner et al., 2009; Wiech and Tracey, 2009) which show that emotional reactive stimulations (e.g., the subliminal perception of emotional pictures or other reactive emotional stimulations) modulate pain perception. A further interesting result which supports this line of reasoning comes from experiments reported in Plassmann et al. (2008), which show, by functional MRI studies, how prices of wine bottles (which represent here a piece of information related to the outcome) can affect the experienced pleasantness of the wine intaking (UCS). Indeed, the experienced pleasantness (which represents here

the UCR) is due to the integration of the active component, x, and the reactive (self-induced) response i_R, which is due to information processing. From the above mentioned results emerges evidence that emotional responses are additive, and can energize or decrease an active stimulation if they share common neuronal populations. Moreover, this property is not limited to emotional responses, but it also holds for other neuronal populations belonging to the reactive system (see Section 2.1). Indeed, pharmacological conditioning experiments (Amanzio and Benedetti, 1999; Benedetti et al., 2011) show that the conditioning (reactive) contribution increases the base active pharmacological effect of a given drug, even in animals (Guo et al., 2010). The additive property of emotional responses which became attributed to a given common stimulus is termed *integration property*. Considering one single neuronal population, the integration property can be expressed as:

$$y = x + i_R, \tag{2}$$

where y represents the experienced CNS response, x represents the active response contribution and i_R the reactive response.

2.3.3. Relation between Predicted and Reactive Response: The Reactive Stability Theorem

As described in the previous sections, whenever a source of stimulation (UCS) elicitates a CNS response (UCR) in a given organism a prediction error is computed as the difference between expected (or predicted) and experienced (i.e., UCR) responses; furthermore such an error signal updates the predicted response (i.e., UCS revaluation). Denoting $y_{predicted}$ and y the predicted and the experienced response respectively, the prediction error computation can be expressed as $e = y - y_{predicted}$. Furthermore, whenever a source of stimulation is perceived by an organism a reactive emotional response has to be elicited (in particular, as shown in the previous sections this is performed by the amygdala). Considering one single component (i.e., a specific neuronal population) of the given reactive emotional response, what can be said about its intensity? Does the reactive response coincide with the expected (or predicted) response? The following theorem proves that the reactive response associated with a given UCS has to be a fraction (i.e., less than the unity) of the expected response, in order to assure the stability of the emotional system. *The emotional system is said to be stable with respect to a given stimulus if and only if the response elicited by the stimulus does not increase unlimitedly over time.*

> **Theorem**: *Necessary condition for the stability of the emotional system is that the emotional response associated to a given UCS is a fraction of the expected (predicted) response.*

The demonstration of the above theorem is provided in "Supplementary Data."

On the basis of the Reactive Stability Theorem and of the reactive mimicking property, the reactive response of the generic neuronal population can be written as:

$$i_R = \alpha \cdot y_{predicted}, \tag{3}$$

where the term α represents the intensity fraction (or *gain*) of the generic mimicked component belonging to the emotional system (such that $|\alpha| < 1$). Generally speaking, if K represents the number of neural populations involved in the emotional response, a vector of K different values for the reactive gain α exists, in which every component is associated to a single neuronal population. The generic term α is also termed *emotional learning rate* in the following.

2.3.4. Emotional Contrast Effects

In the technical literature it is well documented (Flaherty, 1982; Papini and Dudley, 1997) that surprising reward omissions, that is, the absence or reduction of an expected reward, are accompanied by aversive emotional reactions. On the other hand, surprising increases in the expected reward result in an appetitive emotional reaction. In particular, positive and negative *contrast effects*, arising from unexpected shifts in the obtained reward (whose value is greater or smaller than that previously experienced), depend on the comparison of the sensory property of the present stimulus with information stored in memory (Genn et al., 2004) and lead to an emotional response overshoot or undershoot, which is independent from the absolute value of the real reward. For instance, in Genn et al. (2004) it is shown that rats, in the presence of a shift from 32% to a 4% of the administered sucrose solution, displayed a successive negative contrast (i.e., a *depression effect* Flaherty, 1982) by initiating significantly fewer bouts of licking than control rats maintained on 4% sucrose. Furthermore, no significant increase in the dopamine efflux in the NAcc was observed during the consumption of 4% sucrose by rats that experienced the shift from 32%; on the contrary, the consumption of 4% sucrose by control rats was accompanied by a significant increase in the DA efflux in the NAcc.

The notion that contrast effects can be interpreted in terms of emotional responses is indirectly suggested by the effects of drugs on contrast (Flaherty, 1982). Indeed, experimental data reveal that drugs having anxiolytic effects on humans (e.g., amobarbital, ethanol, and benzodiazepines) tend to reduce negative contrasts. Furthermore, experimental results reviewed in Flaherty (1982) show an increase of the release of adrenocorticosteroid hormones in the presence of negative contrasts; this proves that a negative contrast is able to activate a component of the sympathetic response to stress, which, in turn, determine an emotional response.

Experimental evidence also shows that contrast exhibits an inverse dependence on the *inter trial interval,* denoted T, (i.e., the time interval between two successive stimulation trials) and a direct dependence with the magnitude difference between the preshift and the postshift values. The inter trial interval dependence suggests that modeling this effect should involve continuous time scale evaluations.

On the basis of the above mentioned results it is clear as emotional contrast effects have to be quantitatively taken into account in the model of the emotional response dynamics, since a difference between expected and actual stimulation determines inevitably an adding quantity in the final CNS response.

2.4. Model Assumptions and Hypothesis

In this Section the main hypothesis and assumptions adopted for the model development are summarized, on the basis of the results reviewed on the previous Sections.

H1 - *Definitions*: the definitions illustrated in Section 2.1 are adopted.

H2 - *Linearity hypothesis*: it is assumed that, at a *system level,* the linearity hypothesis holds for error signals and responses.

H3 - *Single emotional component*: we focus on the dynamics of a *single component* to ease the reading. This choice, however, does not entail any loss of generality, since our model can be applied to any component of the emotional system.

H4 - *Integration property*: as illustrated in Section 2.3.2 reactive and active responses add up as in Equation (2).

H5 - *Functional connectivity*: the structure of the dynamical emotional system (both in the discrete and in the continuous time scales) is expressed in **Figure 1**.

H6 - *Learning mechanisms*: it is assumed that both type of learning (CC and UCS revaluation) can co-occur simultaneously, and they are subjected to the constraints derived in Section 2.2 (see **Figure 2**).

H7 - *Prediction error computation*: prediction errors are computed as the difference between the expected/predicted and the experienced responses; furthermore, we will consider two different hypothesis for the expected response: (a) it coincides with the experienced response in the last trial, (b) it is computed as a filtered version (i.e., a weighted moving average) of the last trials outcomes.

H8 - *Stability of the emotional system property* holds (see Section 2.3.3 and Equation 3).

H9 - *Source Attribution*: it is assumed only one eliciting UCS and that it is correctly attributed by the emotional system. When a different scenario has to be considered it will be specified.

H10 - *Emotional contrast effects*: negative and positive contrast effects are evaluated as a linear function of the discrepancy between the expected and the incoming outcome.

H11 - *Discrete trials* (valid in the discrete time scale) - Multiple trials in the interaction between a source and a subject are considered; the trial duration $\triangle T$ is assumed to be relatively small and, in particular, negligible with respect to the *inter-trial interval* (ITI) T. For this reason, each trial can be ideally associated with a specific point on the time axis and the corresponding emotional response can be deemed constant.

H12 - *Residual response from previous trials* (valid in the discrete scale) - The time constant τ associated with the decay of the response elicited during each trial is deemed negligible respect to the inter-trial interval T; consequently, when a new trial takes place, the emotional response due to the previous trials has already vanished.

H13 - *Stimulus (UCS) perception* - It is assumed that the perceived UCS is the same in each trial, so as the associated contextual information and boundary conditions. This assumption states that, if a stimulus elicits a subject during the first trial in a specific context (e.g., place, timing, and specific boundary conditions), it has to be considered that the *stimulus perception* in the following trials involves exactly the same contextual and boundary conditions. In absence of such an assumption the reactive response elicited by the stimulus perception might be modulated by the different contextual information and boundary conditions. For instance, if an UCS is represented by a given drug, which has been encoded as UCS because previous interactions, then "UCS perception" refers to the UCS intake (in order to satisfy the same conditions occurred during the previous UCS-subject interactions), so that the reactive response associated with such a UCS can be triggered (independently from that the active pharmacological treatment has been altered).

H14 - *Recurrent patterns of stimulation* - A source of stimulation can elicit an organism with some regularities over time (or over discrete trials). For instance an electric shock device could stimulate a subject performing a periodic intensity pattern over time, or a given drug can exert a specific pattern of active effect over time (e.g., pharmacodynamic curve-related effects).

2.5. Model Development

2.5.1. Discrete Time UCS Revaluation Model (Without Conditioning)

Motivation: the model accounts for a given UCS eliciting an organism with a variable active stimulation (x) and/or a variable reactive stimulation (i_R).

Hypothesis: H1-5, H7a, H8-9, H11-13.

The discrete model is obtained through a *thought experiment* in which a given subject is stimulated by an UCS over successive trials. More specifically, the target UCS is perceived by the subject at every trial, in order to induce a reactive stimulation, after that an active UCS stimulation follows (e.g., through an electric shock delivery).

Provided that multiple stimulation trials are considered, it can be assumed that in every trial the expected (or predicted) response associated with the given UCS coincides with the last experienced outcome (which, in turn, coincides with the response experienced in the previous trial, H7a), or, alternatively, that the predicted response converges over successive trials to the actual experienced outcome (H7b). Without any loss of the generality, and with a first order approximation, it can be assumed that the predicted outcome is equal to the last experienced outcome. The expected response is updated through the prediction error computation over successive trials, in turn, the reactive response i_R will be updated according to Equation (3) . In the first trial the reactive response is equal to zero, since the emotional system did not have any past learning experiences or interactions with the given UCS (so that the epxected response is zero); hence the elicited response is exclusively determined by the active stimulation:

$$y_1 = x_1. \tag{4}$$

Since the expected outcome was equal to zero for that UCS, the prediction error after the first active stimulation trial is equal to x_1 and, such an error updates the expected response for the new trial, and, consequently the reactive response, according to Equation (3), that is:

$$y_{expected,\ 2} = x_1 \tag{5}$$

$$i_{R2} = \alpha \cdot x_1. \tag{6}$$

In the second trial, as soon as the UCS is perceived the learned reactive response will be triggered, which, together with the active stimulation determine the CNS response (see Equation 2):

$$y_2 = x_2 + \alpha x_1; \tag{7}$$

moreover, a new error signal is computed as

$$e_2 = y_2 - y_1 = x_2 + \alpha x_1 - x_1. \tag{8}$$

Without loss of the generality, it can be assumed that the active elicitation is kept constant at every trial (i.e., $x_n = x\ \forall n$, where n represent the trial index), so that the error at the second trial can be expressed as

$$e_2 = \alpha \cdot x \tag{9}$$

and the reactive response updated at the end of the second trial is given by

$$i_{R3} = \alpha x + \alpha^2 x. \tag{10}$$

It easy to demonstrate that the response elicited in the n-th trial can be expressed as:

$$\begin{array}{c} y_n = x_n + \alpha \cdot \sum_{k=1}^{n-1} e_k = \\ x_n + \alpha \cdot \sum_{k=1}^{n-1} (y_k - y_{k-1}), \end{array} \tag{11}$$

which, can be reformulated, as:

$$y_n = x_n + \alpha \cdot y_{n-1}. \tag{12}$$

The last equation shows that the overall CNS response is determined by the contributions of the active elicitation (x_n) and of the reactive (emotional) contribution, determined by previous learning, and expressed as a fraction of the expected outcome, which, it has to remember, it is assumed to coincides with the response elicitation in the previous trial, with a first order approximation.

If a constant active elicitation x is considered over successive trials, it is easy to show that the response approaches the asymptotic value

$$y_\infty = \frac{x}{1-\alpha} \tag{13}$$

as n increases.

When the asymptotic value has been reached, the prediction error will be zero for every successive stimulation trials, and no predicted response updating can occur. The error signal will be zero also if the condition $y_n = y_{n-1}$ occurs in the generic n-th trial. As will be shown in the Section "Results," some psychophysiological phenomena (e.g., evaluative conditioning) and neuropsychiatric pathologies (e.g., PTSD) occur if the above mentioned condition holds (together with the following conditions: (a) the reactive response is different from zero and (b) the active response is zero. Both conditions can be expressed in terms of the following: the expected or predicted response coincides with the reactive/self induced response). Furthermore, it is easy to prove that if a series of successive trials in which the active stimulation (x) is kept equal to zero, the CNS response in the n-th trial can be expressed as:

$$y_n = \alpha \cdot y_{n-1} = y_0 \alpha^n, \tag{14}$$

where y_0 represents the expected response before the beginning of the UCS devaluation process. Hence, during devaluation, the response tends asymptotically to zero.

Contrast Effects: hypothesis H10 is added in the model.

Contrast effects can be included in the discrete model of implicit emotional learning by adding a new function, called *contrast function* and denoted C_{eA}, which, generally speaking, could be a function of the *actual error-signal* (denoted e_A), defined as

$$e_{A,n} \triangleq \left(x_n + \alpha \cdot y_{n-1}\right) - y_{n-1} \tag{15}$$

for the n-th trial; note that this definition is motivated by the fact that the actual error signal refers to the actual trial (instead of the previous one), since contrast effects occur in parallel with the actual outcome. Hence, the emotional response in the n-th trial can be expressed as (see also Equation 12)

$$y_n = x_n + \alpha \cdot y_{n-1} + C_{eA} \cdot e_{A,n} \tag{16}$$

if $e_{A,n} \neq 0$ and

$$y_n = y_{n-1} \tag{17}$$

if $e_n = 0$ and $e_{A,n} = 0$.

Assuming a simple linear contrast function $C_{eA} \simeq K$ and assuming $0 < K < 1$ (for emotional stability reasons), it is easy to demonstrate that Equation 12 becomes

$$y_n = (1+K) \cdot x_n + (\alpha + K\alpha - K) \cdot y_{n-1}. \tag{18}$$

On the basis of the above reported results, if an unexpected UCS active elicitation occurs (i.e., an active UCS stimulation which is not signaled by any CS nor by a prior UCS perception; for instance, this scenario can be represented by a laboratory setting where a permanently-connected electric shock device elicits a subject without any prior signaling), it determines the response

$$y_{UCS} = x + K \cdot x, \tag{19}$$

and is attributed to the UCS. Moreover, a prediction error is computed and the reactive response associated with the UCS is updated; more specifically, if the expected response before the unexpected elicitation was equal to $x + \alpha x$, the prediction error is computed as $e = x \cdot (K - \alpha)$. Furthermore, if another unexpected UCS elicitation occurs, the resulting prediction error is equal to zero since the expected outcome is now equal to the actually experienced outcome, which is given by $x + K \cdot x$ (i.e., is determined by the active elicitation x and the contrast contribution due to the unexpected elicitation Kx). This mathematical result shows that a series of trials of unexpected

UCS elicitations lead to a computation of a prediction error only in the first unexpected elicitation; indeed, in the successive unexpected trials only a constant reactive contribution (i.e., *Kx*) due to the contrast effect is elicited. Indeed, if at every unexpected UCS stimulation an error signal was computed, then the UCS revaluation would lead to an unbounded increase of the UCS expected response, which actually is not the case. The above mentioned results and observations lead to a novel interpretation of experimental results (Schultz et al., 1997; Hollerman and Schultz, 1998; Schultz, 2002) in relation to the recording of the activity of dopamine neurons in the VTA during unexpected rewarding stimulations (See Section 3.2.4). Hence, unexpected stimulations represent a particular case of emotional contrast effect, in which the expected response was zero; furthermore, since no error signal is computed, such an increase in dopamine response due to unexpected stimulation simply reflect a reactive response, which, in turn, may subserve as an incentive to orienting and focus attention on the source of stimulation and on eventual suspicious statistical or contextual contingencies (in other words, in this case the dopamine response is not computed to update the UCS value, but for focus attention in order to observe if some contingent cues with the unexpected UCS release occurred). It is worth noting that the same situation occurs also when an expected reward is omitted (i.e., negative contrast effect), in this case the induced negative reactive response does not update UCS values, instead it focuses attention to discover eventual contingent cues (such cues, if do exist, can become *conditioned inhibitors*; Harris et al., 2014). As will be shown in Section 3.2.4 dopamine neurons in the VTA and substantia nigra can subserve to other brain regions (e.g., the OFC) to compute both error signals and reactive responses (associated to UCSs, or due to contrast effects in order to focus attention and facilitate further learning).

Hypothesis, H7b: the expected response is computed as an exponential weighted average of the last trials outcomes. This hypothesis is motivated by the consideration that the expected/predicted responses can be shaped considering different previous outcomes and not only the last one. This hypothesis is confirmed by experimental results (Bayer and Glimcher, 2005) which show that dopamine neurons *encode the difference between the current reward and an exponentially weighted average of previous rewards.*

Under the H7b hypothesis, the predicted response, denoted $< y_{n-1} >$ (since it represents the filtering function of the last responses until that occurred in the n-1 trial), can be expressed as:

$$< y_{n-1} >= \sum_{k=1}^{L} h_k \cdot y_{n-k}, \tag{20}$$

where L represents the number of the responses involved in the filtering process, and h_k represents the generic k-th filter coefficient (or weight). Note that Equation (20) expresses a general moving filtering function and not only an exponential one. It can be shown that a moving exponentially averaged filter (in discrete time) represents a *low pass filter* (and also that its continuous-time counterpart is an R-C first order type filter; Schafer and Oppenheim, 2009). Substituting Equation (20) in Equation (12) yields:

$$y_n = x_n + \alpha \cdot \sum_{k=1}^{L} h_k \cdot y_{n-k}. \tag{21}$$

Furthermore, considering that the error signal is now computed as:

$$e_k = y_k - < y_{k-1} >, \tag{22}$$

Equation (21) can also be written as (see also the first line of Equation 11):

$$y_n = x_n + \alpha \cdot \sum_{k=1}^{n-1} \left(y_k - \sum_{v=1}^{L} h_v \cdot y_{n-v} \right). \tag{23}$$

Adding Hypothesis, H14: the active stimulation presents a recurrent pattern over successive trials (e.g., a gaussian shaped curve for the intensity of the stimulation occurs periodically).

If the brain recognizes recurrent patterns of stimulation over time (i.e., a typical stimulation intensity pattern), such an information will be exploited for a more precise inference of the probable outcome. This line of reasoning is supported by experimental results which show that neurons encode precise timings between stimuli (Schultz et al., 1997); for instance, dopamine neurons learn, after few observations, that after a prescribed time from the onset of a cue a given quantity of reward will be delivered.

Learning statistical regularities and patterns represent a type of CC learning, since the variable "time" and "temporal relations" can be considered as contextual cues (Bouton, 1993). For this reason we speculate that the pattern recognition function may be performed by the hippocampus, together with OFC interactions for the inference of more complex patterns. In practice, what is learned in this case is that at a given "time reference" (CS1) a specific UCS intensity stimulation occurs, then at a successive time reference (CS2) a different UCS intensity stimulation occurs and so on, until an eventual entire recurrent pattern of stimulation will be learned. From a modeling perspective it is important to note that if the brain "is sure" about the fact that a specific pattern of stimulation is occurring (denoted $y_{P(1..N)}$ where the interval $(1..N)$ represents the entire range of trials the pattern comprises), then the predicted response at the generic n-th trial is represented by the corresponding intensity within the pattern (i.e., y_{Pn}). Conversely, if the brain does not recognize any pattern, no adding inference can be performed and the predicted response is computed as in Equation (20). Nevertheless, intermediate situations between the above mentioned extremes can occur; more precisely during pattern learning, or whenever the recognized probability of having a given pattern is not equal to one, the expected response has to be computed as a combination of the actual UCS revaluation contribution (Equation 20) and of the response

expected by the pattern. It is easy to prove that the predicted (expected) response can be expressed as:

$$< y_{n-1} >= \omega_{NP}^{(n)} \cdot \sum_{k=1}^{L} h_k \cdot y_{n-k} + \omega_P^{(n)} \cdot y_{Pn}, \quad (24)$$

where $\omega_P^{(n)}$ ($\omega_{NP}^{(n)}$) represent the weight (or *belief confidence*) related to the occurrence (no occurrence) of the given pattern at the n-th trial; furthermore, it has to be satisfied the following equality:

$$\omega_{NP}^{(n)} = 1 - \omega_P^{(n)}. \quad (25)$$

Equation (24) shows that the predicted response comprises two contributions: (a) the UCS revaluation component, which represents the *bottom-up contribution,* since it is determined by the actual perception of the response (or its gradient over time) and it is exploited from higher level neuronal networks to form more complex hierarchical patterns; (b) the contribution due to previous inferential learning, which represents the *top-down contribution*, since it is encoded by higher level hierarchical neural structures and exploited to compute a reactive response which will be perceived by lower level structures for the computation of the prediction error.

In order to include the contrast effect in the model an additional term proportional to the *actual error signal* has to be considered. It is easy to show that the actual error signal can be computed as:

$$e_{A,n} = x_n + \left(\omega_{NP}^{(n)} \sum_{k=1}^{L} h_k \cdot y_{n-k} + \omega_P^{(n)} y_{Pn} \right) \cdot (\alpha - 1). \quad (26)$$

Hence, the discrete emotional learning model of a periodic stimulation of period N_0 (i.e., such that $y_n = y_{n-N_0}$) can be written as:

$$y_n = x_n + \alpha \omega_{NP}^{(n)} \sum_{k=1}^{L} h_k \cdot y_{n-k} + \alpha \omega_p^{(n)} y_{n-N_0} + K \cdot e_{A,n} \quad (27)$$

and its implementation is reported in **Figure 3**.

2.5.2. Discrete Time Classical Conditioning Model with Implicit UCS Revaluation

Motivation: derivation of a discrete model for CC under the stochastic Hebbian plasticity hypothesis and considering the implicit UCS revaluation during acquisition process.

Hypothesis: H1-6, H7a, H8-13.

In this Section a discrete model of classical conditioning which accounts for the implicit UCS revaluation is presented, and its derivation is developed through a thought experiment, where a sequence of trials, involving CS-UCS-subject interactions, is analyzed. Our derivation relies on the assumption that the CS-UCS synaptic connections are governed by the mechanisms of *stochastic Hebbian plasticity* (Hebb, 1949; Amit and Fusi, 1994; Fusi, 2002; Soltani and Wang, 2006, 2010; Fusi and Abbott, 2007). This hypothesis is supported by both some experimental results shown in Redondo et al. (2014) and Gore et al. (2015), and other models relying on the fact that a CS-UCS pairing entails the Hebbian potentiation of the CS inputs onto the UCS representations in the BLA (Sah et al., 2003; Pickens and Holland, 2004; Pape and Pare, 2010). Hebbian learning is based on the idea that synapses between neurons being simultaneously active become stronger. Consequently, "neurons that fire together wire together" through an increase in synaptic efficacy mediated by *long term potentiation* (LTP); on the other hand, a decrease in synaptic efficacy is mediated by *long term depression* (LTD). The mathematical derivation of the proposed model is available in "Supplementary Material" Section.

The mathematical analysis results in the following model for classical conditioning in the discrete time scale:

$$\omega_{CS-UCS}^{(n)} = \omega_{CS-UCS}^{(n-1)} + \hat{\alpha}_+ \cdot \left(1 - \omega_{CS-UCS}^{(n-1)}\right) \quad (28)$$

$$\omega_{CS-UCS}^{(n)} = \omega_{CS-UCS}^{(n-1)} - \hat{\alpha}_- \cdot \omega_{CS-UCS}^{(n-1)} \quad (29)$$

$$i_R^{(n)} = \alpha \cdot \left(X + i_R^{(n-1)} \cdot \omega_{CS-UCS}^{(n)}\right) \quad (30)$$

$$y_{CS}^{(n)} = \omega_{CS-UCS}^{(n)} \cdot i_R^{(n-1)}. \quad (31)$$

Note that Equations (28) and (31) hold for $n \geq 2$ and that the initial conditions

$$\omega^{(1)} = 0 \quad (32)$$

$$i_R^{(1)} = \alpha \cdot X \quad (33)$$

$$y_{CS}^{(1)} = 0 \quad (34)$$

and

$$y_{UCS}^{(1)} = X \quad (35)$$

should be adopted when employing them. The term $\omega_{CS-UCS}^{(n)}$ is termed *synaptic strength* and represents the fraction of synapses from the neurons representing the CS stimulus onto the encoding neurons for the UCS in the n-th trial; the terms $\hat{\alpha}_+$ and $\hat{\alpha}_-$ in Equations (28 and 29) represent the *potentiation* and the *depression* rates respectively, and they determine the probability for plastic synapses to switch from the depressed to the potentiated state and vice-versa. If UCS revaluation during conditioning is neglected the proposed model coincides with the well known Rescorla-Wagner (R-W) model (Miller et al., 1995). Our extended model provides a more general and accurate description of the emotional response during conditioning than the original R-W model for different reasons which are described in Section 3.2.4.

2.5.3. Conditioning to a Reactive Source Stimulus

Three possible events can occur when a CS is conditioned to a purely reactive UCS (i.e., an UCS for which no active component elicitation is expected, e.g., an emotional picture): (1) a simple associative connection CS-UCS is generated; (2) the CS is misattributed to be the source of the elicited response, so becoming a new (and independent) reactive source like the original UCS, so that a new reactive response, equal (but independent) to the original i_R, is generated and associated to the CS. Furthermore, this misattribution process could occur even during the presentation of the CS alone after conditioning, since the reactive elicitation (which is equal to $\omega_{CS-UCS} \cdot i_R$) could be misattributed forward the CS, in this case the reactive response being associated with CS corresponds to the quantity $\omega_{CS-UCS} \cdot i_R$. (3) A combination of the previous two events could occur. Moreover, if one of the two last mentioned events occurs, the conditioned CS may become an inextinguishable element of emotional reaction, since the expected response is purely reactive and coincides with the elicited reactive response (i.e., the prediction error is always zero).

2.6. Emotional System Dynamics in Continuous Time Scale

In the previous Sections a discrete-time model for the computation of the emotional response in different scenarios has been developed. In real world conditions, however, a stimulation might elicit continuously a subject. From a modeling perspective a continuous elicitation can be seen as a series of an infinite number of discrete trial stimulations, each of which has an infinitesimal time duration and the temporal spacing between them tends to zero. In these conditions the emotional response is continuously updated driven by the continuous time counterpart of the prediction error. In the following, the problem of developing a system computational model for describing the continuous time evolution of the emotional response is investigated. More specifically, our approach is based on standard engineering methods. Without any loss of the generality we focus here on the development of the model in absence of the pattern recognition contribution, since we focus here on the dynamics of the emotional control system; furthermore, such a feature can be modeled by a neural network for pattern recognition which operates in parallel with the OFC.

In order to obtain a continuous counterpart of a discrete model the *sampling time* has to be known. Generally speaking the sampling time is defined as the time period at which the continuous time model is sampled to obtain a discrete version of it. In our analysis we started developing directly a discrete model, since the experimental results available in the literature are based on discrete trials measures. For this reason we have now to infer the continuous model whose sampling would produce the discrete model obtained in the previous sections. Considering that the discrete model holds for any arbitrary large inter-trial interval (denoted T) and that we are interested in finding the emotional dynamics in the limit of T which tends to zero, it is possible to

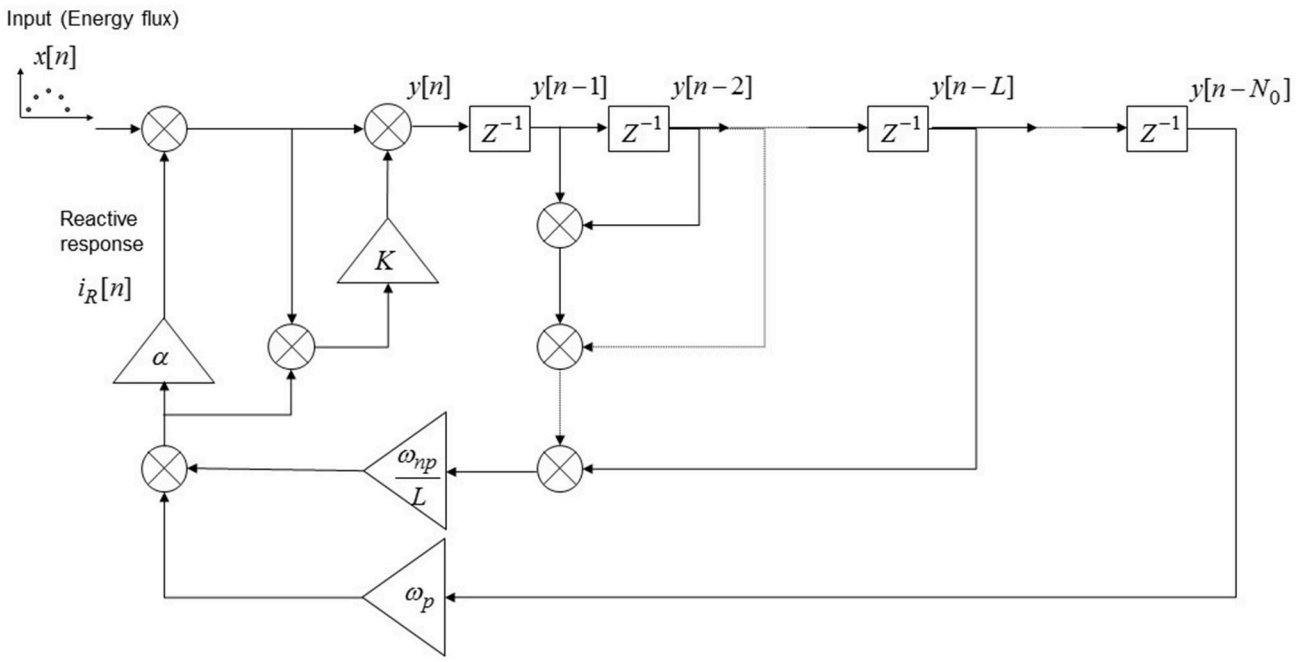

FIGURE 3 | Discrete-time model implementation of the emotional system dynamics. The input to the system ($x[n]$) is represented by a series of discrete stimulations over successive trials; the emotional response in the n-th trial is given by $i_R[n]$ and the overall CNS response is given by $y[n]$. The blocks "Z^{-1}" represents one unit delay, the triangular blocks represent multiplicative factors and all the nodes are summation nodes. The model takes into account of the implicit UCS revaluation, of the contrast effect and of the "pattern recognition" in the case of a periodical stimulation pattern of period N_0.

consider the smallest T such that the discrete model holds, and then assuming it as the sampling time. The *sensory time discrimination threshold* (Luna et al., 2005) represents the smallest temporal interval for which the CNS neurons can discriminate between two distinct consecutive stimulations, hence, this parameter is taken as the sampling time (T). It is worth pointing out that the value of T depends on the involved perceptive modality (e.g., somatosensory stimulation, visual stimulation or acoustic stimulation), and that, if an active stimulation varies faster than T the neurons encode an average value within a time window of T. For instance, providing that in the visual stimulation T is about 100 ms, the stimuli variation within 100 ms are encoded as a mean value over 100 ms.

Assuming the discrete model can be obtained sampling the continuous time counterpart of it with a sampling time T, the continuous model can be obtained through the following procedure:

(a) the *transfer function* of the discrete recursive difference system in the Z-domain (Schafer and Oppenheim, 2009) is computed; (b) the transfer function of the continuous dynamic system is obtained in the S-Laplace domain substituting the Z variable of the discrete transfer function with the following equation:

$$Z = \frac{1 + s \cdot T/2}{1 - s \cdot T/2} \tag{36}$$

where the variable s represents the Laplace variable (this substitution represents the inverse operation of the so called *bilinear transform*; Schafer and Oppenheim, 2009); (c) if needed, the differential equation in the time domain (or the continuous time state space representation) is obtained with the inverse Laplace Transform of the equation in the Laplace domain obtained in the previous stage b.

Applying the aforementioned procedure to the discrete model (Equation 27) the continuous dynamic system can be developed and its *transfer function* in the Laplace domain can be expressed as:

$$\begin{aligned} H(s) &= \frac{Y(s)}{X(s)} \\ &= \frac{s^2 + K\tau_2 + s\,(K + \tau_2) + 1}{s^2\tau_1\,(\tau_2 + K - K\alpha) + s\,(\tau_2 + K - K\alpha + \tau_1 - \alpha\tau_1) + 1 - \alpha} \end{aligned} \tag{37}$$

where the term τ_1 represents the time constant of the equivalent low pass filter relative to the $x(t)$ neuronal population target (and it is closely related to the time discrimination threshold T); the term τ_2 represents the equivalent time constant of the low pass filtering effect performed by the emotional evaluation system (i.e., the equivalent of the exponential weight moving average in the discrete time model); $Y(s)$ and $X(s)$ represent the overall response and the input (i.e., the active stimulation) in the Laplace domain respectively. Taking into consideration the constraints and the functional connections derived in the previous sections (see **Figure 1**), the implementation of the derived transfer function model can be obtained as depicted in **Figure 4**.

It is easy to prove that the system differential equation obtained from the continuous model represented in **Figure 4** is:

$$\begin{aligned} \ddot{y}(t) \cdot \tau_1\,(\tau_2 + K - K\alpha) + \dot{y}(t)\,(\tau_2 + K - K\alpha + \tau_1 - \alpha\tau_1) \\ + y(t)\,(1 - \alpha) = \ddot{x}(t) \cdot K\tau_2 + \dot{x}(t)\,(K + \tau_2) + x(t) \end{aligned} \tag{38}$$

where the functions $y(t)$, $\dot{y}(t)$, $\ddot{y}(t)$,$x(t)$, $\dot{x}(t)$ and $\ddot{x}(t)$ represent the elicited response, its first and second derivatives, the active stimulation and its first and second derivatives over time, respectively.

3. RESULTS

In this Section some results from the theory and the developed models, understood as postpredictions and quantitative explanations of experimental observations reviewed from the technical literature, are presented. Successively a model comparison with existing models is provided. Furthermore, a section describing model validation, interpretation and applicability to some research topics is presented.

3.1. Summary of the Derived Models

All the models are based on the functional structure described in **Figures 1**, **2**. The main assumptions are summarized in Section 2.4.

A simplified discrete model for UCS revaluation and contrast effects is described by Equation (18) and it is named M1 in the following.

The discrete model for UCS revaluation, contrast effects and pattern recognition (named M2) is expressed by Equation (27).

The discrete model for CC and implicit UCS revaluation (named M3) is expressed by Equations (28–35).

The continuous time system dynamical model (named M4) can be expressed by the transfer function in Equation (37) and represented in **Figure 4**, and by Equation (38).

3.2. Post-predictions and Quantitative Explanations

3.2.1. Resistant-to-Extinction Responses

Generally speaking, on the basis of the theory and models derived in the previous sections, an emotional response can become resistant to extinction (or inextinguishable) if (1) the prediction error is zero while (2) the reactive response is different from zero while (3) the active response is zero; or, equivalently, if the reactive response coincides with the expected response. Indeed, since, in general, the emotional system tracks an active component, it is obvious that whenever such a component decrease or vanishes then the associated reactive response will be decreased too. However, in the continuous time scale, there are two cases which can lead to a situation in which the active stimulation drops to zero and the reactive response does not (so that it becomes inextinguishable): (a) if a saturation level of the expected response is reached; (b) in the presence of specific dynamics of the active stimulation, exploiting the inertial nature of the emotional tracking system. We will describe in detail the

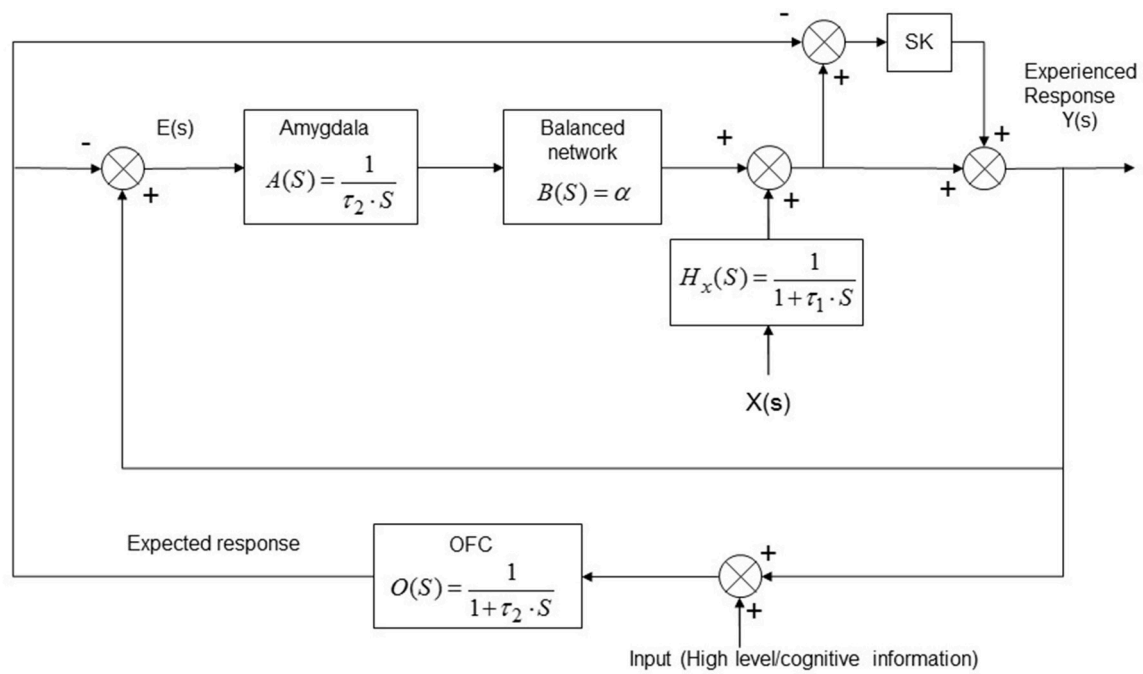

FIGURE 4 | Continuous-time model implementation of the emotional system. The model is presented in the Laplace domain where "S" represents the Laplace operator. The functional connections between the involved actors and the processing task accomplished by each block are shown. The model takes into account of the implicit UCS revaluation and of the contrast effect. X(s) represents the input (energy flux), Y(s) the CNS response and E(s) represents the prediction error; $\alpha$is the emotional learning rate for the neuronal target population, K is the contrast effect parameter. OFC, orbitofrontal cortex.

first case, and we propose future computational and experimental researches to investigate the second.

On the other hand, in the discrete time scale it is possible inducing resistant to extinction responses through the misattribution phenomena (see Section 2.2.6). In fact, if a purely reactive response is attributed (and hence expected) to a given stimulus, then it will induce the same reactive response which is expected (i.e., the prediction error is always zero). A quantitative analysis of such a case will be provided next.

3.2.2. Inextinguishability through Emotional Response Saturation and Hippocampus Impairment

During an extreme traumatic event different phenomena may occur, and, our model shows how these conditions can determine a resistant-to-extinction emotional reaction.

Generally speaking, a traumatic response is mainly determined by an automatic inferential emotional learning, instead that by an active energy-based stimulation *per se*. More specifically, the processing of traumatic information, such as an imminent death danger, can be implicitly processed by the OFC which, in turn, activates the amygdala through the computation of prediction errors. Indeed, experimental results from the literature (Steinberg et al., 2013; Sadacca et al., 2016) show that *inferred* outcomes (e.g., rewards), never directly experienced before, determine prediction errors (for instance computed in the VTA) which are just like predictions based upon direct experience/stimulation. In practice this means that the OFC can control the computation of error signals, which stimulate the amygdala, on the basis of the *difference of the actual expected outcome and the inferred outcome* (through information processing) associated with a given UCS or situation, even before experiencing a direct active (energy-based) stimulation. From a modeling perspective this situation can be described assuming that the *input* to the emotional system is the *error node* (see **Figure 4**), and that the prediction error is proportional to the difference between the inferred and the expected outcome. Indeed, experimental results based on optogenetic manipulations (Chang et al., 2016) have shown that inducing artificial prediction errors within the VTA permits to induce behaviors and responses like those obtained by inference and statistical learning. Moreover, it has been observed that prediction errors are transmitted in spikes-form (i.e., very short and relatively small decrease in neuron firing rates); we argue this modality represents a suitable method to send direct (inferred) error signals since the transfer function between the *error node* (E(s)) and the output (Y(s)) is unstable, so that a continuous error signal could rapidly lead to an unlimitedly increasing of the output (it is easy to prove that the Y(s)/E(s) transfer function represents an unstable system; see **Figure 4**).

Nevertheless, if the inferred outcome is relatively greater than the expected response encoded within the OFC (associated with the target situation or stimulation), then the inferred prediction error could become so intense to determine a sort of saturation of the amygdala response. Indeed, any mathematical function

$f(x)$ representing a specific *biological response* cannot take on arbitrarily large values, because of the limited dynamics of the response itself. In practice, as the value taken on by the argument x grows, the corresponding value $f(x)$ of the function does not steadily increase in proportion to it and, when x crosses a certain threshold, a certain saturation level is reached; in other words, $f(x)$ exhibits a *nonlinear* behavior for sufficiently large values of x. In particular, these considerations hold for the amygdala, which takes prediction errors as input and it updates the predicted outcome on the basis of such errors; we denoted such a behavior $F_A(e) = Y_{expected}$ (within the amygdala). Hence, the linearity hypothesis holds if in any trial (or at any specific time instant) the prediction error takes on a value smaller than a *saturation threshold* T_S, which defines the linearization range for the amygdala expected response function ($F_A(e)$; see **Figure 5**). If, however, the prediction error exceeds T_S (i.e., if the error becomes excessively large), a phenomenon of *emotional saturation* should be expected. When emotional saturation occurs, the source of stimulation (or any associated cue) generating it could produce inextinguishable effects, even in the absence of any active stimulation or if the exposition to a trauma-related stimulus occurs in a *safe context*. In fact, if the prediction error computed because a safe context is smaller than the reached degree of saturation, then its inhibitory effect does not determine any response reduction(see **Figure 5**). Moreover, it is important to note that contextual information are primarily coded and stored in the hippocampus (while the representation of an aversive stimulus is coded within the BLA), so that, if during the traumatic event the hippocampus does not properly code contextual information, then no effective contextual discrimination can be obtained during further exposures of the stimulus. As far as this last point is concerned, it is worth mentioning that hippocampus functioning and its ability to encode contextual information are impaired by uncontrollable stress together with an hyperactivation of the amygdala, as shown in a model developed in Kim and Diamond (2002). Furthermore, if hippocampus activities are impaired during high stress exposure, it is likely that the amygdala associates every contextual cue directly (even if insignificant) with the eliciting reactive response (i.e., misattribution), so that, successively, they will be able to trigger the reactive response (Bechara et al., 1995). These considerations lead us to the conclusion that the traumatic emotional response becomes resistant-to-extinction since the expected response coincides with the reactive triggered response. We argue that this phenomenon could happen in panic disorders and PTSD (Beck and Sloan, 2012; Parsons and Ressler, 2013; Perusini et al., 2016). As a matter of fact, in some forms of PTSD and panic disorders the mere repetitive exposure to cues related to a traumatic event does not lead to an extinction of emotional responses or results in a very slow extinction (Paunovic, 1999; Van Rooij et al., 2015; Perusini et al., 2016).

Finally, it is important pointing out that the standard classical conditioning model is unable able to explain these psychopathologies, since it does not account for UCS revaluation nor for the conditions which lead to the prediction error to be equal to zero while an emotional response greater than zero occurs.

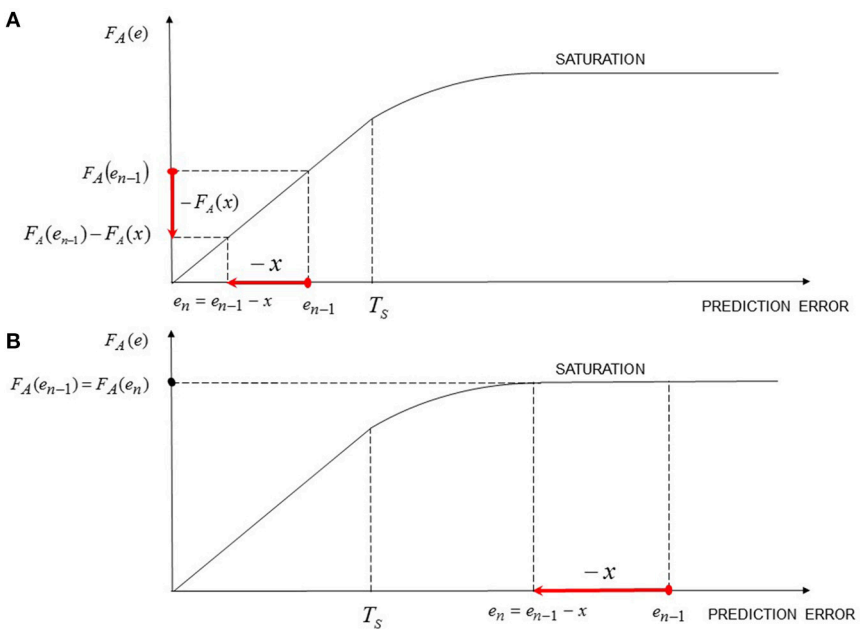

FIGURE 5 | Resistant-to-extinction emotional reaction through response saturation. Schematic representation of the biological behavior of the amygdala function $F_A(e)$ in its linear zone **(A)** and in its saturation zone **(B)**. In case **(A)** the prediction error (equal to $-x$) is able to reduce the elicited emotional response. On the contrary, in case **(B)** a negative prediction errors unable to produce a similar effect, so that the emotional response remains at its saturation level. In particular, the case **(B)** occurs if the negative prediction error (due, for instance, to the fact that the active response x is no more elicited during the stimulation or because the stimulation is occurring in a safe context) is smaller than the degree of saturation reached by the amygdala in the previous stimulation(s).

Figure 6 show results of simulations performed assuming a traumatic and a non-traumatic stimulation. In the former the amygdala response saturation and hippocampus impairment occurred, since the prediction errors due to the difference between the inferred and the expected responses are relatively intense (e.g., death danger). In order to simulate a non-traumatic scenario we have assumed that the expected outcome within the OFC was greater than in the traumatic case, so that the resulting inferred error signals are smaller. Otherwise, it can be assumed that the prediction errors are smaller because the stimulation is less stressful. Nevertheless, assuming a given stressful stimulation, the model suggests that if the expected response within the OFC (i.e., the conscious expectation) is high, the reached response intensity under the stimulation is lower than in the case of a lower expected outcome, and the further extinction is facilitated.

3.2.3. Inextinguishability Due to Misattribution of a Reactive Source of Emotional Stimulation

In order to show how a reactive misattribution leads to a resistant-to-extinction response let us consider the model M1 without H9, and focus in a thought experiment in which multiple trials with interaction between a source of stimulation and a subject occur. Furthermore, let us assume that the elicited response is misattributed to another source of stimulation. In the following we assume, without any loss of generality, that the misattributed source of stimulation is initially *neutral* (i.e., it does not elicit an active or a reactive emotional response). However, if the misattributed source of stimulation is not neutral but elicits a response, the response elicited during the misattribution process will result from the superposition of the actual source response with the previous non-attributed emotional state (Zillmann, 1971). For this reason, in this case the previous non-attributed emotional state "energizes" the actual source.

The emotional misattribution process encompasses the following three mutually exclusive cases:

(a) Misattribution occurs in the presence of an active stimulation.
(b) Misattribution occurs in the presence of a residual (i.e., passive) response decay only (in other words, no active or reactive responses are elicited), in this case the misattribution trial follows the elicitation trial and occurs during the excitation decay. This case is known in literature as *transfer paradigm* (described in the Hullian *drive theory* Hull, 1943) or also as *excitation transfer* (Zillmann, 1971; Zillmann et al., 1972; Bunce et al., 1993), and refers to the influence of a prior episode of arousal on subsequent emotional responses.
(c) Misattribution occurs when a purely reactive source of stimulation is eliciting the subject, so that the associated response is purely reactive.

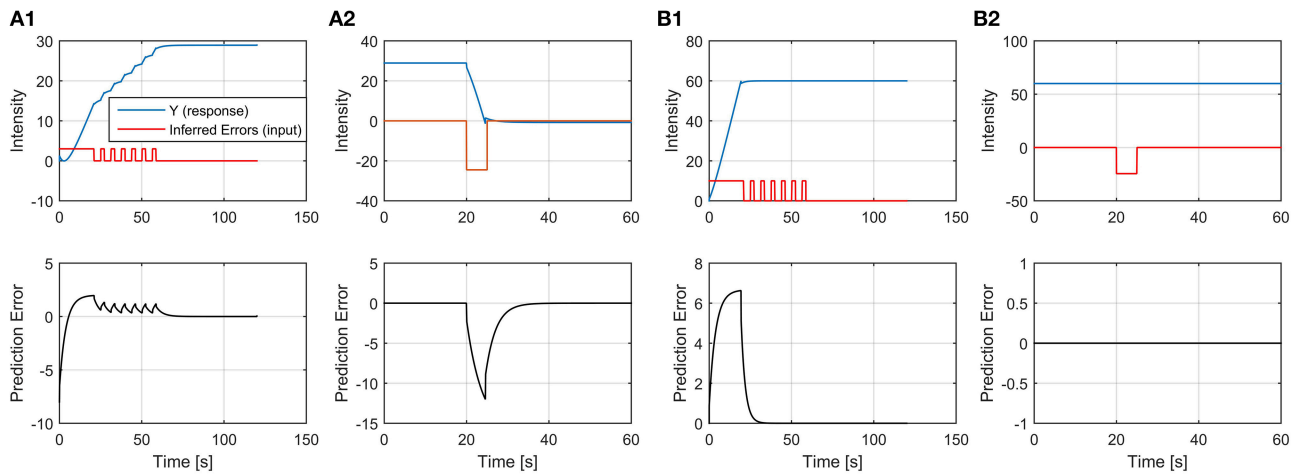

FIGURE 6 | Simulations of emotional responses determined by prediction errors based on inference learning. (A1) A stressful (non-traumatic) reactive response acquisition: the OFC determines a train of inferred prediction errors which are sent to the amygdala, based on information flux processing. More specifically, the subject consciously expected a negative outcome (i.e., Y expected within the OFC is greater than zero) and the inferred negative outcome is not extreme or traumatic, so that no saturation level of the emotional response, nor hippocampus impairment occur. **(A2)** Physiological devaluation of the previously learned response: the subject is exposed to a stressful-related cue but in a safe context; the inhibition (i.e., deflation) of the emotional reaction determines the devaluation of the response in the safe context. **(B1)** Traumatic response acquisition: inferred prediction errors generate a response which grows at relatively high level, causing both amygdala response saturation and hippocampus impairment. In this case the subject does not expect such an intense negative outcome (i.e., Y expected within the OFC is negligible with respect to the inferred outcome) and/or the inferred outcome is extremely intense. **(B2)** The traumatic response cannot be devaluated since the inhibitory inferred prediction error due to the detection of a safe context is not sufficient to determine a response reduction. This can be determined by two main phenomena: (1) the hippocampus impairment during the traumatic event blocked the encoding of the contextual information, moreover, for this reason, some contextual stimuli have been misattributed as causal sources of stimulation and coded within the amygdala; (2) the weak inhibition due to the safe context is smaller than the degree of saturation reached by the amygdala during the trauma exposure. (*Prediction error* patterns represent the time course of the difference between the expected (within the OFC) and the experienced (within the CNS) responses). Parameter values: model M4, $\alpha = 0.4$; $K = 0.5$; $\tau = 2$s; saturation level for $Y_{expected}(amygdala) = 150$ (which determines a reactive response saturation equals to 60); the level of OFC expectation before the stimulation is equal to 8 in **(A1)** and to 1 in **(B1)**; the inferred outcome is equal to 10 in both **(A1, B1)**.

In this last cited case, of our interest here, the response in the first trial (when the misattribution is occurring), denoted i_U, is due to a purely reactive response elicited by an unrevealed or confused source (e.g., a subliminal emotional picture stimulation Esteves et al., 1994; Mayer and Merckelbach, 1999; Glascher and Adolphs, 2003). Hence, the response attributed to the new source (because of the misattribution) is equal to i_U, which, in this case, coincides with the response that the amygdala is already eliciting within the CNS. More specifically, since i_U represents the CNS response eliciting by the amygdala, the intensity of the amygdala response (and of the expected response) can be expressed as i_U/α (see **Figures 1**, **2** and Equation 3). Furthermore, the prediction error in the first trial, which is equal to i_U, is sent to the amygdala, and, hence, the overall expected response encoded within the BLA will be due to the sum of the prediction error and the response already elicited by the amygdala, that is $i_U/\alpha + i_U$. Hence, during the second trial, the exposure to the new source determines the reactive response $\alpha\,(i_U/\alpha + i_U)$ which is greater than the expected value previously stored within OFC (that was i_U), this leads to the computation of a new prediction error equal to $\alpha \cdot i_U$.

Following this line of reasoning it is easy to prove that the reactive response can be expressed

$$y_n = i_U + i_U \sum_{k=1}^{n-1} \alpha^k \tag{39}$$

after n exposure trials, so that the reactive response asymptotically converges to

$$y_\infty = \frac{i_U}{1-\alpha}. \tag{40}$$

Nevertheless, as it is shown is Section 2.2.6 an UCS can be attributed only partially through a statistical inference; this means that in real situations only a portion of the entire response is attributed to the new misattributed UCS, and that successive misattribution trials can improve the attributed response. After some trials in which an emotional picture (UCS1) is paired with a previously neutral picture (CS) the misattribution process, when occurs (since it is stochastic), leads to an overall reactive (and expected) response associated to CS which is comprises between 0 and $i_U/(1-\alpha)$ (see Equation 40).

A concrete example of such an effect is the EC through *implicit misattribution* (Jones et al., 2009; Hutter and Sweldens, 2013).

Evaluative Conditioning (EC) represents the formation (or change) of the valence of a stimulus, called CS, originating from a prior pairing of the CS itself with another stimulus, called UCS (Baeyens et al., 1993; De Houwer et al., 2001; Jones et al., 2009; Hofmann et al., 2010; Gast et al., 2012; Hutter and Sweldens, 2013); unlike Pavlovian conditioning, a CS response acquired through EC seems to be resistant to extinction (Baeyens et al., 2005).

In Jones et al. (2009) it is shown that, according to the implicit misattribution model, responses to UCSs can be misattributed without awareness to the CS, and that the implicit misattribution depends on *source confusability*. More specifically, the subject may confuse which multiple occurring stimuli in her environment is evoking the evaluative response.

Furthermore, manipulations of the variables related to the potential for the misattribution of an evaluation, (i.e., the source confusability) show that greater EC occurs with an higher degree of confusability (Jones et al., 2009).

This result is also supported by Baeyens et al. (1993), who found that EC was not sensitive to the degree of statistical contingency between the CS and the UCS (as happens in classical conditioning), but EC should increased with the absolute number of pairings because each provides an opportunity for misattribution, and such misattributions could act cumulatively (Jones et al., 2009).

In conclusion, our model based on prediction errors explains, from a quantitative perspective, the EC phenomena and post-predicts the above cited results.

3.2.4. Predictions and Results of the Discrete Model

3.2.4.1. Classical conditioning

M3 model predicts all the relevant results predicted by R-W model for CC (see Miller et al., 1995), since, starting from stochastic hebbian plasticity hypothesis for CS-UCS synaptic connection, the model results in an extended version of the R-W model where, in addiction, accounts for the implicit UCS revaluation. In other words, neglecting UCS revaluation M3 coincides with the R-W model. Moreover, M3 is able to quantitatively justify some experimental results not predictable with R-W model, such as *the dependence of asymptotic responding on CS intensity and US intensity* (Young et al., 1976; see also Miller et al., 1995 and articles therein). Indeed, Equations (30 and 31) show that the CR intensity influences the intensity of the unconditioned response, since, even if the changes of i_R over successive trials could be really small, the asymptotic value of i_R is $\alpha X/(1-\alpha)$, which is greater than the initial value αX. This leads to the conclusion that, since the value of the parameter α is influenced by the selected CS (if the impact of other factors, such as internal physiological states and the selected UCS, is deemed constant), different CSs may result in distinct asymptotic values of i_R (and consequently of y_{CS}; see Equation 31).

Moreover, M3 provides a more general and accurate description of the emotional response during conditioning than the original R-W model for different reasons. First of all, it includes the contributions of both the active response (X) due to the elicitation of the UCS and the reactive (self-induced) response associated with the UCS representation within the BLA (i_R). Moreover, the recursive equations describing it are *causal* unlike those representing the R-W model. Note that causality ensures that the currently computed response depends only on the past and present values of the stimulus and the response itself, but not on their future values; unluckily, this does not occur for the Rescorla-Wagner model since the evaluation of the current response requires the knowledge of the final asymptotic response, which is actually unknown to the brain. Finally, in our model the CS-UCS synaptic strength and the consequent UCS inflation are jointly considered: the model shows that classical conditioning learning influences the reactive response associated

with the paired UCS; this is due to the misattribution of the CR contribution forward the UCS response.

Furthermore, we argue that the model (M1-3) motivates also the *spontaneous recovery* (Miller et al., 1995) and the *conditioned inhibition and all the related phenomena* (such as the *failure of the extinction of conditioned inhibition through non-reinforced presentations of the inhibitor*; DeVito and Fowler, 1987; Harris et al., 2014), which cannot be described in terms of classical conditioning nor TD models (the mathematical demonstrations are not reported in this manuscript).

3.2.4.2. Dopamine neurons activity predictions and relation between existing models

The derived discrete model (M1-3) explain the experimental measures on the activity of dopamine neurons within the VTA (see Schultz, 2002 and articles therein). Indeed, (1) before CC learning a rewarding UCS perception elicits a reactive response; (2) during CC acquisition dopamine neurons respond progressively to the onset of the CS and no more to the UCS; (3) if an UCS stimulation occurs unexpectedly the dopamine neurons activity is physically increased; (4) dopamine neurons encoded the difference between the current reward and an exponentially weighted average of previous rewards (Bayer and Glimcher, 2005). Our model is able to justify such dopamine neurons behaviors, more specifically, whenever the CS-UCS connection strength increase the reactive response (i_R) associated with the given UCS can be activated by the CS perception through the synaptic connection ω_{CS-UCS} (see **Figure 2**; nevertheless, before any cue association, only the UCS perception is able to activate the reactive response (all the intermediate situations in which $0 < \omega_{CS-UCS} < 1$ are also predicted). Moreover, if successive reward stimulations occur the error signal at each trial is effectively computed as an exponentially weighted average of previous outcomes; furthermore whenever an unexpected reward occurs an adding reactive response due to contrast effect is triggered. Such a contrast reactive response determines an error signal which update the UCS expected value only the first time it occurs (see Section 2.5.1), and this assures that an unbounded increasing of the UCS associated value does not occur. The last reasoning supports the idea that the reactive contrast effect is different than prediction error and it represents a type of reactive response. Indeed, other than reward prediction errors, different types of dopamine neural coding exists (Bromberg-Martin et al., 2010). More specifically, dopamine neurons can in some cases compute prediction errors, and in other cases code or compute reactive responses associated with given stimuli (CSs or UCSs) in order to promote attention and specific (approaching or avoiding) behaviors. The same function is performed by contrast effects which have to focus attention toward the stimulator promoting further learning (for instance, discovering a specific cue which may in the future predict such an unexpected stimulation). One of the difference between our theory and TD models for reinforcement learning (Schultz et al., 1997; Schultz, 1998) is that the latter assume that dopamine neurons encode only prediction errors (and does not take into consideration nor define the term "reactive response"), and that learning can occur only in the presence of a prediction error. Nonetheless, there are evidences which disagree with TD models: for instance, the occurrence of associative learning between two neutral stimuli (the so called, *sensory pre-conditioning;* Young et al., 1998; Sadacca et al., 2016) or the evidence obtained by fMRI studies in which the conditioned acquisition, or, in other words, an increase of the CS-UCS contingency occurs even in presence of a negative prediction error due to the concurrent deflation (decrease) of the UCR. In practice, if during acquisition (CS-UCS pairings) the intensity of the UCS stimulation is reduced, the CS-UCS connection is still increased but a negative prediction error is computed and updates the expected outcome associated to the UCS (Gottfried and Dolan, 2004). Also the opposite situation may occur: during conditioning extinction a positive prediction error due to UCS inflation can be obtained. Such discrepancies originate from the fact that TD (and R-W) models do not account for the two main different types of learning (CC and, more generally, the *statistical and inferential learning* and *UCS revaluation driven by prediction errors*, see **Figure 1**). In particular, the inferential and statistical learning creates a so called *model of the world* (Doll et al., 2012), and it is not driven by errors but it occurs through statistical (e.g., Bayesian) inference. Such an inference about the statistical contingencies and causalities can be performed by the hippocampus (and even directly by the amygdala) for low level information processing, or by OFC whenever complex pattern or higher level information have to be analyzed (see **Figure 1**). Furthermore, as we have previously shown (see Equation 24), the statistical inference contribution and the prediction error based contribution (also called *model-free* contribution; Doll et al., 2012) are taken into consideration by the brain for the computation of the expected response depending on the degree of certainty (belief) of each of the two components.

In conclusion, our model post-predicts the most relevant phenomena related to learning and dopamine neurons, moreover it predicts further important related phenomena with respect to existing learning models.

3.3. Validation, Interpretation, and Applicability of the Model

3.3.1. Validation of the Discrete Model

The discrete model parameters (i.e., α,K, and the filtering coefficient, together with the estimation of the induced x) can be estimated, for every emotional component, inducing specific stimulation trials while neuronal activity is monitored (e.g., by fMRI or direct neurons activity recording). For instance increasing intensity electric shock delivery can be performed estimating the parameters valid for fear and anxiety related emotional responses, and from unexpected stimulations the contrast parameter$K$ can be estimated; furthermore, rewarding stimulations can be induced in order to estimate the parameters valid for the dopamine neuron populations (e.g., in VTA).

3.3.2. Discrete Model Applications

The discrete model can be adopted in different psychological paradigms and experiments other than the study of dopamine neurons behavior. For instance, it is well known from the literature (Bechara et al., 1994) that patients with damaged OFC (and PFC) perform poorly at the Iowa Gambling Task (IGT).

It is thought that such patients cannot learn from previous emotional error signals, and, in line with this reasoning the structure derived in **Figure 1** shows that if OFC is damaged the emotional prediction error computation is compromised. Our model can be applied to the study of IGT, more precisely, a given deck of cards represents an UCS whose average stimulation has to be estimated, and the single cards represent specific emotional stimulations (in this case purely reactive, since no active/energy based stimulations occur). Firstly the model parameters have to be inferred by successive stimulations and neuronal activity measurements (e.g., by fMRI). Such parameters are: α, K, h_k(which represents a unique exponential coefficient considering an exponential weight average filtering), $x_i(i = 1, ...R)$ where R represents the number of the possible card outcomes and x_i represents the emotional response associated to the i-th stimulation card (e.g., a gain of 50$). Such parameters have to be estimated monitoring the neuronal activity while, from a unique deck, successive cards (with random order) are discovered by the subject (for instance, the subject perceive the sequence +50$, +100$, -50$...). Once known the parameters related to a given subject (or to a group of subjects), the model can predict the performance of that subject at IGT, provided that the sequence pattern (i.e., cards sequences) are given. More specifically, some patients can perform poorly because an altered K parameter, others because an altered α parameter, or because an unbalanced reactive response associated with positive (rewarding) with respect to the negative cards (x_i), and so on. For instance similar studies have been proposed for patients with Parkinson's disease (Zaghloul et al., 2009).

3.3.3. Validation of the Continuous Time Model

M4 can be validated in different scenarios. For instance the time constant τ_2 related to the filtering process and the parameterα (for a target population) can be estimated applying a (constant) direct neuronal stimulation to the target population and measuring the overall response over time (this could be accomplished adopting optogenetic manipulation technology (Redondo et al., 2014), performing the so called *step response measure*). In this scenario the asymptotic value is related to α and to the direct stimulation by the relation seen in Equation (13), while the rise time is related to τ_2. Furthermore, the model can be tested applying a time varying stimulation over time while recording the overall response. Finally, even a time varying function representing the error signal dynamics can be directly applied to the neuronal population involved in error signal computations (e.g., the VTA) while the overall response dynamics is recorded in the target population (e.g., in the NAcc), similarly as performed in Chang et al. (2016).

3.3.4. Applications and Modulation (Increasing/Decreasing) of Emotional Responses

The derived model can be adopted for the study of the emotional dynamics during a continuous stimulation. A practical example is music, in which complex hierarchical patterns of acoustic sound successions (which involve inferential learning/pattern recognition) together with continuous modulation of contrast effects are exploited to evoke specific emotional responses. It is well known that music is able to evoke emotions, for instance, violating expectations or shifting in time the rewards in a balanced mechanism based on frustration (i.e., tension, as a state of dissonance, instability and uncertainty Huston et al., 2015) and satisfaction (resolution toward consonant and stable sounds experienced as pleasurable; Koelsch, 2014). Violation or retardation in resolution produces a tension increase which may result in a successive stronger satisfaction during resolution (Huston et al., 2015). However, it has to remember that the specific mapping function between the features of a given physical source, which drives the energy flux, and the corresponding active emotional response induced by it (i.e., $x(t)$ or x_n, which represents the mean firing rates of a given neuronal population elicited by the energy flux) has to be determined. If this function is known, the physical features of the source can be controlled in a way to generate specific dynamics in the active response; this, in turn, results in the generation of designed emotional reactive responses.

3.3.4.1. Artificial emotional modulation and production of resistant-to-extinction responses

Our theoretical findings suggest that the "inertial nature" of the *emotional dynamic tracking system* can be exploited to originate a resistant-to-extinction emotional reaction; this, in turn, may be exploited to increase (decrease) an emotional response until a saturation level (zero). In the following is explained how this can be obtained.

Starting from Equation (38) it is possible performing numerical optimization procedures to find $x(t)$ patterns which determine a CNS response (denoted $y(t)$) such that an $y(t)$ different from zero occurs while a "very close to zero" prediction error ($e(t)$) occurs together with the condition that $x(t)$ is close to zero, within a sufficient long time interval. The functional to be optimized has to involve all the above mentioned conditions. Despite the fact the so obtained reactive response could be relatively small, the stimulus which is associated to such a pattern of stimulation will acquire a resistant-to-extinction response; furthermore, since an emotional reaction (even if small) is triggered, even in absence of an active stimulation (since it is resistant-to-extinction) the presentation of a new input stimulation function (i.e., $x(t)$), with the same dynamics of the previous one, will permit to obtain an increasing of the response, exploiting a summation effect (see the *integration property* and Equation 2). Hence, with a limited (and periodical) dynamics of the input $x(t)$ is possible to obtain an "unlimitedly" increasing emotional reaction exploiting the inertial nature of the emotional system. In order to test the hypothesis, after having obtained the desired $x(t)$ function from numerical optimizations, such a function can be applied to CNS emotional population (e.g., the anterior cingulate regions, or rewarding brain regions such as NAcc) through optogenetic manipulations (Redondo et al., 2014), or through direct electric neuronal stimulation, while recordings the increase of the overall population emotional response over time. We argue that at every application of the optimized $x(t)$ there is a probability greater than zero that the "inextinguishability effect" takes place, adding a contribution

to the previously accumulated resistant-to-extinction response.

3.3.5. Inducing Traumatic (Saturated) Responses: Testing the Model

On the basis of the analysis and models developed in the previous sections, we argue it is possible to induce traumatic emotional responses in a laboratory through optogenetic manipulations. More specifically, the misattribution effect and the implicit UCS revaluation can be exploited in an iterative framework until a saturation level will be reached. It is proposed the following procedure: (1) the animal has to infer that a given electric shock device (UCS1) is the causal source of pain; in particular, at every stimulation trial UCS1 has to be perceived by the animal, emitting also a brief specific acoustic tone (CS1, which will serve only as "probe" to test the response inextinguishability at the end of the inductive process) just before the occurring of the electric shock stimulation. It is important to note that every time that UCS1 is presented to the animal, it will elicit the electric shock. (2) Some stimulation trials have to occur in order to reach the asymptotic response $y = x/(1-\alpha)$, provided that x represents the active component. (3) A second pain stimulator device (UCS2), different from UCS1 (for instance a device producing pain by heat shock can be adopted) has to be presented to the animal exactly as UCS1. (4) Neuronal representations (memory engrams) of UCS1 and UCS2 have to be detected and labeled within the BLA. At this point it is important to note that two distinct reactive responses, i_{R_1}, i_{R_2}(such that $i_{R_1} = \alpha x_1/(1-\alpha)$; $i_{R_2} = \alpha x_2/(1-\alpha)$), have been associated with the BLA memory engrams of UCS1 and UCS2 respectively, so that it is possible to activate such reactive responses by optical stimulation of the associated engram cells (Ramirez et al., 2015). (5) In successive UCS1 stimulation trials, the BLA memory engram associated with UCS2 has to be optically stimulated in order to induce an UCS1 revaluation determined by the sum of the active (electric shock elivery, denoted x), the UCS1 reactive response i_{R_1} and the reactive response associated with UCS2, i_{R_2}. The above mentioned revaluation occurs since the overall response (active and reactive) will be fully attributed to UCS1 (in other words, a misattribution does occur). (6) Some stimulation trials have to be performed in order to reach the new asymptotic response attributed to UCS1, which will be equal to:

$$y_{UCS1}^{1-\infty} = x_1 + i_{R_1} + \frac{i_{R_2}}{1-\alpha} = x_1 + i_{R_1} + \frac{\alpha i_{R_2}}{1-\alpha} + i_{R_2} \qquad (41)$$

where, $y_{UCS1}^{1-\infty}$ represents the asymptotic UCR1 at the end of the first iterative procedure; (7) The UCS2 stimulates the animal, while optical stimulation of the UCS1 memory engram occurs, so that the UCR2 in the first trial of the second iterative procedure can be expressed as:

$$y_{UCS2}^{2-1} = x_2 + i_{R_2} + i_{R_1}. \qquad (42)$$

It is worth noting that, at this stage, the reactive response i_{R_1} has been increased during the first iterative procedure, and its value was increased from $\alpha x_1/(1-\alpha)$ to (see Equation 41):

$$i_{R_1}^{1-\infty} = \frac{\alpha x_1}{1-\alpha} + \frac{\alpha i_{R_2}}{1-\alpha} = \frac{\alpha x_1}{1-\alpha} + \frac{\alpha^2 x}{(1-\alpha)^2}. \qquad (43)$$

If, without any loss of generality, it is assumed that $x_1 = x_2 = x$ and that $\alpha = 0.5$ in order to simplify the computations, the UCR2 asymptotic value at the end of the second iterative procedure can be expressed as:

$$y_{UCS2}^{2-\infty} = x + i_{R_2} + \frac{i_{R_1}^{1-\infty}}{1-\alpha} = x + x + 4x = 6x \qquad (44)$$

(8) procedures 6 and 7 are repeated iteratively, increasing i_{R_1} and i_{R_2}at every stimulation trial; we named this procedure *iterative climbing*, since the derived protocol resembles a climbing performed by leaning iteratively between UCS1 and UCS2. It is easy to verify by induction that the process leads to a response which diverge to infinity (i.e., $y_{UCS1}^{\infty-\infty} \to \infty$). In practice, it is expected that when a saturation threshold is reached, the error signal will be zero and no more response increases can occur. In such a situation the emotional response will be resistant-to-extinction and the simple presentation of the UCS1 with the probe CS1 (without active stimulation) to the animal will produce the traumatic reaction. Indeed, UCS1 (and CS1) represent the traumatic triggering cues; it is also expected that the only CS1 presentation is able to trigger such a traumatic response like in PTSD patients.

4. DISCUSSION

In this manuscript a system computational model of emotional learning has been developed. The model shows the differentiation (and the relations) between statistical inference learning (e.g., CC) and implicit UCS revaluation, and provides various new insights on well known psychophysiological phenomena and psychiatric diseases, and new ideas for further research. One of its most interesting implications is represented by the identification of well defined mathematical and neurophysiological conditions ensuring the inextinguishability of specific emotional reactions. In particular, it allows us to establish the following four different mechanisms through which a stimulus can produce a resistant-to-extinction emotional reaction: (1) misattribution of a reactive response; (2) classical conditioning of a stimulus to a purely reactive UCS; (3) saturation of emotional response together with hippocampus impairment (also reproducible through optogenetic manipulations exploiting the *iterative climbing procedure*); (4) the exploitation of the inertial dynamics of emotional system on a continuous time scale. Further relevant contributions are represented by the proof that the Rescorla-Wagner model for classical conditioning can be obtained as a special case of the proposed model; the derivation of a new model for conditioning, which accounts for the implicit UCS revaluation and that is able to quantitatively describe important experimental results, which are unpredictable by existing classical conditioning models (including the TD model).

Our result paves the way for various new research activities. First of all, various potential applications of our theory can

be envisaged in the hot research area concerning the study (and the manipulation) of animal behaviors, emotional reactions and decision making, since our model permits to infer specific parameters involved in emotional induced responses under decision, which are known to influence (or even drive) human decisions (see Bechara et al., 1994).

A further relevant research topic concerns the applications of our model of emotional learning on a continuous time scale. Generally speaking, this model could be exploited to analyze the emotional reaction generated by any stimulation which varies continuously over time (e.g., a time-varying acoustic source of stimulation, such as music, or even a purely reactive emotional induction, such as a succession of emotional pictures, or a movie).

Finally, it is important to mention that our theoretical framework can be exploited for the development of animal psychophysiological experimental models; these, in turn, can potentially provide new insights into emotion-related phenomena and pathologies.

AUTHOR CONTRIBUTIONS

LP and SR wrote the main manuscript text; SR prepared **Figures 1–6**. All authors reviewed the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fncom.2016.00054

REFERENCES

Amano, T., Duvarci, S., Popa, D., and Pare, D. (2011). The fear circuit revisited: contributions of the basal amygdala nuclei to conditioned fear. *J. Neurosci.* 31, 15481–15489. doi: 10.1523/JNEUROSCI.3410-11.2011

Amanzio, M., and Benedetti, F. (1999). Neuropharmacological dissection of placebo analgesia: expectation-activated opioid systems versus conditioning-activated specific subsystems. *J. Neurosci.* 19, 484–494.

Amit, D., and Fusi, S. (1994). Dynamic learning in neural networks with material synapses. *Neural Comput.* 6, 957–982. doi: 10.1162/neco.1994.6.5.957

Anderson, C. (1989). Temperature and aggression: ubiquitous effects of heat on occurrence of human violence. *Psychol. Bull.* 106, 74–96. doi: 10.1037/0033-2909.106.1.74

Baeyens, F., Diaz, E., and Ruiz, G. (2005). Resistance to extinction of human evaluative conditioning using a between-subjects design. *Cogn. Emot.* 19, 245–268. doi: 10.1080/02699930441000300

Baeyens, F., Hermans, D., and Eelen, P. (1993). The role of cs-us contingency in human evaluative conditioning. *Behav. Res. Ther.* 31, 731–737. doi: 10.1016/0005-7967(93)90003-D

Bayer, H. M., and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47, 129–141. doi: 10.1016/j.neuron.2005.05.020

Bechara, A., Damasio, A. R., Damasio, H., and Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 7–15. doi: 10.1016/0010-0277(94)90018-3

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., and Damasio, A. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* 269, 1115–1118. doi: 10.1126/science.7652558

Beck, J., and Sloan, D. (2012). *The Oxford Handbook of Traumatic Stress Disorders*. New York, NY: Oxford University Press.

Benedetti, F. (2008). Mechanisms of placebo and placebo-related effects across diseases and treatments. *Annu. Rev. Pharmacol. Toxicol.* 48, 33–60. doi: 10.1146/annurev.pharmtox.48.113006.094711

Benedetti, F., Carlino, E., and Pollo, A. (2011). Hidden administration of drugs. *Clin. Pharmacol. Ther.* 90, 651–661. doi: 10.1038/clpt.2011.206

Benedetti, F., Pollo, A., Lopiano, L., Lanotte, M., Vighetti, S., and Rainero, I. (2003). Conscious expectation and unconscious conditioning in analgesic, motor, and hormonal placebo/nocebo responses. *J. Neurosci.* 23, 4315–4323.

Berns, G., McClure, S., Pagnoni, G., and Montague, P. (2001). Predictability modulates human brain response to reward. *J. Neurosci.* 21, 2793–2798.

Bourdy, R., and Barrot, M. (2012). A new control center for dopaminergic systems: pulling the VTA by the tail. *Trends Neurosci.* 35, 681–690. doi: 10.1016/j.tins.2012.06.007

Bouton, M. (1993). Context, time and memory retrieval in the interference paradigms of pavlovian learning. *Psychol. Bull.* 114, 80–99. doi: 10.1037/0033-2909.114.1.80

Bray, S., and O'Doherty, J. (2007). Neural coding of reward-prediction error signals during classical conditioning with attractive faces. *J. Neurophysiol.* 97, 3036–3045.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron* 68, 815–834. doi: 10.1016/j.neuron.2010.11.022

Bryant, J. (2003). *Communication and Emotion: Essays in Honor of Dolf Zillmann, New Edn.* Routledge Communication Series (Routledge), 39–40.

Bunce, S., Larsen, R., and Cruz, M. (1993). Individual differences in the excitation transfer effect. *Pers. Indiv. Diff.* 15, 507–514. doi: 10.1016/0191-8869(93)90333-X

Cacioppo, J. T., Tassinary, L. G., and Berntson, G. (2007). *Handbook of Psychophysiology*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511546396

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H. J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nat. Neurosci.* 19, 111–116. doi: 10.1038/nn.4191

Choi, J., and Jeansok, J. (2010). Amygdala regulates risk of predation in rats foraging in a dynamic fear environment. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21773–21777. doi: 10.1073/pnas.1010079108

Colloca, L. (2014). Emotional modulation of placebo analgesia. *Pain* 155, 651. doi: 10.1016/j.pain.2014.01.009

Cotton, J. (1981). A review of research on schachter's theory of emotion and the misattribution of arousal. *Eur. J. Soc. Psychol.* 11, 365–397.

Davey, G. (1989). Ucs revaluation and conditioning models of acquired fears. *Behav. Res. Ther.* 27, 521–528. doi: 10.1016/0005-7967(89)90086-7

De Houwer, J., Thomas, S., and Baeyens, F. (2001). Associative learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning. *Psychol. Bull.* 127, 853–869. doi: 10.1037/0033-2909.127.6.853

De la Fuente-Fernandez, R., Ruth, T. J., Sossi, V., Schulzer, M., Calne, D., and Stoessl, A. (2001). Expectation and dopamine release: mechanism of the placebo effect in parkinson's disease. *Science* 293, 1164–1166. doi: 10.1126/science.1060937

De la Fuente-Fernandez, R., and Stoessl, A. (2002). The placebo effect in parkinson's disease. *Trends Neurosci.* 25, 302–306. doi: 10.1016/S0166-2236(02)02181-1

De Pascalis, V., Chiaradia, C., and Carotenuto, E. (2002). The contribution of suggestibility and expectation to placebo analgesia phenomenon in an experimental setting. *Pain* 96, 393–402. doi: 10.1016/S0304-3959(01)00485-7

Delgado, M., Li, J., Schiller, D., and Phelps, E. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 363, 3787–3800. doi: 10.1098/rstb.2008.0161

Deneve, S., and Machens, C. K. (2016). Efficient codes and balanced networks. *Nat. Neurosci.* 19, 375–382. doi: 10.1038/nn.4243

DeVito, P., and Fowler, H. (1987). Enhancement of conditioned inhibition via an extinction treatment. *Anim. Learn. Behav.* 15, 448–454. doi: 10.3758/BF03205055

Doiron, B., and Litwin-Kumar, A. (2014). Balanced neural architecture and the idling brain. *Front. Comput. Neurosci.* 8:56. doi: 10.3389/fncom.2014.00056

Dolan, R. J. (2007). The human amygdala and orbital prefrontal cortex in behavioural regulation. *Philos. Trans. R. Soc. Lond. B* 362, 787–799. doi: 10.1098/rstb.2007.2088

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* 22, 1075–1081. doi: 10.1016/j.conb.2012.08.003

Eippert, F., Bingel, U., Schoell, E., Yacubian, J., Klinger, R., Lorenz, J., et al. (2009). Activation of the opioidergic descending pain control system underlies placebo analgesia. *Neuron* 63, 533–543. doi: 10.1016/j.neuron.2009.07.014

Enck, P., Benedetti, F., and Schedlowski, M. (2008). New insights into the placebo and nocebo responses. *Neuron* 59, 195–206. doi: 10.1016/j.neuron.2008.06.030

Esteves, F., Parra, C., Dimberg, U., and Ohman, A. (1994). Nonconscious associative learning: pavlovian conditioning of skin conductance responses to masked fear-relevant facial stimuli. *Psychophysiology* 31, 375–385. doi: 10.1111/j.1469-8986.1994.tb02446.x

Fanselow, M., and Poulos, A. (2005). The neuroscience of mammalian associative learning. *Annu. Rev. Psychol.* 56, 207–234. doi: 10.1146/annurev.psych.56.091103.070213

Flaherty, C. (1982). Incentive contrast: a review of behavioral changes following shifts in reward. *Anim. Learn. Behav.* 10, 409–440. doi: 10.3758/BF03212282

Flykt, A., Esteves, F., and Ohman, A. (2007). Skin conductance responses to masked conditioned stimuli: phylogenetic/ontogenetic factors versus direction of threat? *Biol. Psychol.* 74, 328–336. doi: 10.1016/j.biopsycho.2006.08.004

Friston, K. (2003). Learning and inference in the brain. *Neural Netw.* 16, 1325–1352. doi: 10.1016/j.neunet.2003.06.005

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi: 10.1371/journal.pcbi.1000211

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE* 4:e6421. doi: 10.1371/journal.pone.0006421

Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biol. Cybern.* 102, 227–260. doi: 10.1007/s00422-010-0364-z

Fusi, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. *Biol. Cybern.* 87, 459–470. doi: 10.1007/s00422-002-0356-8

Fusi, S., and Abbott, L. (2007). Limits on the memory storage capacity of bounded synapses. *Nat. Neurosci.* 10, 485–493. doi: 10.1038/nn1859

Gallagher, M., McMahan, R., and Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. *J. Neurosci.* 19, 6610–6614.

Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* 37, 1297–1310. doi: 10.1016/j.neubiorev.2013.03.023

Gast, A., Gawronski, B., and De Houwer, J. (2012). Evaluative conditioning: recent developments and future directions. *Learn. Motiv.* 43, 79–88. doi: 10.1016/j.lmot.2012.06.004

Genn, R., Ahn, S., and Phillips, A. (2004). Attenuated dopamine efflux in the rat nucleus accumbens during successive negative contrast. *Behav. Neurosci.* 118, 869–873. doi: 10.1037/0735-7044.118.4.869

Glascher, J., and Adolphs, R. (2003). Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *J. Neurosci.* 23, 10274–10282.

Goebel, M., Trebst, A., Steiner, J., Xie, Y., Exton, M., Frede, S., et al. (2002). Behavioral conditioning of immunosuppression is possible in humans. *FASEB J.* 16, 1869–1873. doi: 10.1096/fj.02-0389com

Gore, F., Schwartz, E., Brangers, B., Aladi, S., Stujenske, J., Likhtik, E., et al. (2015). Neural representations of unconditioned stimuli in basolateral amygdala mediate innate and learned responses. *Cell* 162, 134–145. doi: 10.1016/j.cell.2015.06.027

Gottfried, J., and Dolan, R. (2004). Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. *Nat. Neurosci.* 7, 1144–1152. doi: 10.1038/nn1314

Gottfried, J. A., O'Doherty, J., and Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301, 1104–1107. doi: 10.1126/science.1087919

Guo, J., Wang, J., and Luo, F. (2010). Dissection of placebo analgesia in mice: the conditions for activation of opioid and non-opioid systems. *J. Psychopharmacol. (Oxford)* 24, 1561–1567. doi: 10.1177/0269881109104848

Haour, F. (2005). Mechanisms of the placebo effect and of conditioning. *Neuroimmunomodulation* 12, 195–200. doi: 10.1159/000085651

Harris, J., Kwok, D., and Andrew, B. (2014). Conditioned inhibition and reinforcement rate. *J. Exp. Psychol.* 40, 335–354. doi: 10.1037/xan0000023

Hebb, D. (1949). *The Organization of Behavior*. New York, NY: Wiley.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., and Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychol. Bull.* 136, 390–421. doi: 10.1037/a0018916

Hollerman, J. R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* 1, 304–309. doi: 10.1038/1124

Hosoba, T., Iwanaga, M., and Seiwa, H. (2001). The effect of UCS inflation and deflation procedures on 'fear' conditioning. *Behav. Res. Ther.* 39, 465–475. doi: 10.1016/S0005-7967(00)00025-5

Hull, C. (1943). *Principles of Behavior: An Introduction to Behavior Theory, The Century psychology series*. New York, NY: D. Appleton-Century Company, Incorporated.

Huston, J., Nadal, M., Mora, F., Agnati, L., and Cela Conde, C. (2015). *Art, Aesthetics, and the Brain*. Oxford: Oxford University Press.

Hutter, M., and Sweldens, S. (2013). Implicit misattribution of evaluative responses: contingency-unaware evaluative conditioning requires simultaneous stimulus presentations. *J. Exp. Psychol. Gen.* 142, 638–643. doi: 10.1037/a0029989

Ito, R., Dalley, J., Howes, S., Robbins, T., and Everitt, B. (2000). Dissociation in conditioned dopamine release in the nucleus accumbens core and shell in response to cocaine cues and during cocaine-seeking behavior in rats. *J. Neurosci.* 20, 7489–7495.

Janak, P., and Tye, K. (2015). From circuits to behaviour in the amygdala. *Nature* 517, 284–292. doi: 10.1038/nature14188

Jensen, K., Kaptchuk, T., Chen, X., Kirsch, I., Ingvar, M., Gollub, R., et al. (2015). A neural mechanism for nonconscious activation of conditioned placebo and nocebo responses. *Cereb. Cortex* 25, 3903–3910. doi: 10.1093/cercor/bhu275

Jensen, K., Kaptchuk, T., Kirsch, I., Raicek, J., Lindstrom, K., Berna, C., et al. (2012). Nonconscious activation of placebo and nocebo pain responses. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15959–15964. doi: 10.1073/pnas.1202056109

Jones, C., Fazio, R., and Olson, M. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *J. Pers. Soc. Psychol.* 96, 933–948. doi: 10.1037/a0014747

Kennerley, S., and Walton, M. (2011). Decision making and reward in frontal cortex: complementary evidence from neurophysiological and neuropsychological studies. *Behav. Neurosci.* 125, 297–317. doi: 10.1037/a0023575

Kim, J., and Diamond, D. (2002). The stressed hippocampus, synaptic plasticity and lost memories. *Nat. Rev. Neurosci.* 3, 453–462. doi: 10.1038/nrn849

Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nat. Rev. Neurosci.* 15, 170–180. doi: 10.1038/nrn3666

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nat. Rev. Neurosci.* 6, 691–702. doi: 10.1038/nrn1747

LeDoux, J. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York, NY: Simon & Schuster.

LeDoux, J. (2000). Emotion circuits in the brain. *Annu. Rev. Neurosci.* 23, 155–184. doi: 10.1146/annurev.neuro.23.1.155

Li, S., and McNally, G. (2014). The conditions that promote fear learning: prediction error and pavlovian fear conditioning. *Neurobiol. Learn. Mem.* 108, 14–21. doi: 10.1016/j.nlm.2013.05.002

Lui, F., Colloca, L., Duzzi, D., Anchisi, D., Benedetti, F., and Porro, C. (2010). Neural bases of conditioned placebo analgesia. *Pain* 151, 816–824. doi: 10.1016/j.pain.2010.09.021

Luna, R., Hernandez, A., Brody, C., and Romo, R. (2005). Neural codes for perceptual discrimination in primary somatosensory cortex. *Nat. Neurosci.* 8, 1210–1219. doi: 10.1038/nn1513

Mayer, B., and Merckelbach, H. (1999). Unconscious processes, subliminal stimulation, and anxiety. *Clin. Psychol. Rev.* 19, 571–590. doi: 10.1016/S0272-7358(98)00060-9

McNally, G., Johansen, J., and Blair, H. (2011). Placing prediction into the fear circuit. *Trends Neurosci.* 34, 283–292. doi: 10.1016/j.tins.2011.03.005

Meuret, A., White, K., Ritz, T., Roth, W., Hofmann, S., and Brown, T. (2006). Panic attack symptom dimensions and their relationship to illness characteristics in panic disorder. *J. Psychiatr. Res.* 40, 520–527. doi: 10.1016/j.jpsychires.2005.09.006

Miller, R., Barnet, R., and Grahame, N. (1995). Assessment of the rescorla-wagner model. *Psychol. Bull.* 117, 363–386. doi: 10.1037/0033-2909.117.3.363

Muramoto, K., Ono, T., Nishijo, H., and Fukuda, M. (1993). Rat amygdaloid neuron responses during auditory discrimination. *Neuroscience* 52, 621–636. doi: 10.1016/0306-4522(93)90411-8

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G., Halbert, S., Wichmann, R., et al. (2015). A circuit mechanism for differentiating positive and negative associations. *Nature* 520, 675–678. doi: 10.1038/nature14366

Nolan, T., Price, D., Caudle, R., Murphy, N., and Neubert, J. (2012). Placebo-induced analgesia in an operant pain model in rats. *Pain* 153, 2009–2016. doi: 10.1016/j.pain.2012.04.026

O'Doherty, J. (2007). Lights, camembert, action! the role of human orbitofrontal cortex in encoding stimuli, rewards, and choices. *Ann. N. Y. Acad. Sci.* 1121, 254–272. doi: 10.1196/annals.1401.036

O'Doherty, J., Dayan, P., Friston, K., Critchley, H., and Dolan, R. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron* 38, 329–337. doi: 10.1016/S0896-6273(03)00169-7

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr. Opin. Neurobiol.* 14, 769–776. doi: 10.1016/j.conb.2004.10.016

Ohman, A. (1993). *Handbook of Emotions*, chapter Fear and anxiety as emotional phenomena. New York, NY: Guilford Press.

Ohman, A. (2005). The role of the amygdala in human fear: automatic detection of threat. *Psychoneuroendocrinology* 30, 953–958. doi: 10.1016/j.psyneuen.2005.03.019

Ohman, A., and Soares, J. (1993). On the automatic nature of phobic fear: conditioned electrodermal responses to masked fear-relevant stimuli. *J. Abnorm. Psychol.* 102, 121–132. doi: 10.1037/0021-843X.102.1.121

Olsson, A., Nearing, K., and Phelps, E. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Soc. Cogn. Affect. Neurosci.* 2, 3–11. doi: 10.1093/scan/nsm005

Olsson, A., and Phelps, E. (2007). Social learning of fear. *Nat. Neurosci.* 10, 1095–1102. doi: 10.1038/nn1968

Pape, H., and Pare, D. (2010). Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. *Physiol. Rev.* 90, 419–463. doi: 10.1152/physrev.00037.2009

Papini, M., and Dudley, R. (1997). Consequences of surprising reward omissions. *Rev. Gen. Psychol.* 1, 175. doi: 10.1037/1089-2680.1.2.175

Parsons, R., and Ressler, K. (2013). Implications of memory modulation for post-traumatic stress and fear disorders. *Nat. Neurosci.* 16, 146–153. doi: 10.1038/nn.3296

Paton, J., Belova, M., Morrison, S., and Salzman, C. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* 439, 865–870. doi: 10.1038/nature04490

Paunovic, N. (1999). Exposure counterconditioning (ec) as a treatment for severe ptsd and depression with an illustrative case. *J. Behav. Ther. Exp. Psychiatry* 30, 105–117. doi: 10.1016/S0005-7916(99)00010-5

Pavlov, I. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. London: Oxford University Press.

Pehlevan, C., and Sompolinsky, H. (2014). Selectivity and sparseness in randomly connected balanced networks. *PLoS ONE* 9:e89992. doi: 10.1371/journal.pone.0089992

Perusini, J. N., Meyer, E. M., Long, V. A., Rau, V., Nocera, N., Avershal, J., et al. (2016). Induction and expression of fear sensitization caused by acute traumatic stress. *Neuropsychopharmacology* 41, 45–57. doi: 10.1038/npp.2015.224

Pessoa, L. (2010). Emotion and cognition and the amygdala: from "what is it?" to "what's to be done?" *Neuropsychologia* 48, 3416–3429. doi: 10.1016/j.neuropsychologia.2010.06.038

Petrovic, P., Kalso, E., Petersson, K., and Ingvar, M. (2002). Placebo and opioid analgesia– imaging a shared neuronal network. *Science* 295, 1737–1740. doi: 10.1126/science.1067176

Pickens, C., and Holland, P. (2004). Conditioning and cognition. *Neurosci. Biobehav. Rev.* 28, 651–661. doi: 10.1016/j.neubiorev.2004.09.003

Plassmann, H., O'Doherty, J., Shiv, B., and Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1050–1054. doi: 10.1073/pnas.0706929105

Puviani, L., and Rama, S. (2016). Placebo response is driven by UCS revaluation: evidence, neurophysiological consequences and a quantitative model. *arXiv:1602.00258*.

Puviani, L., Rama, S., and Vitetta, G. M. (2016). Prediction errors drive UCS revaluation and not classical conditioning: evidence and neurophysiological consequences. arXiv:1601.07766v1

Ramirez, S., Liu, X., MacDonald, C., Moffa, A., Zhou, J., Redondo, R., et al. (2015). Activating positive memory engrams suppresses depression-like behaviour. *Nature* 522, 335–339. doi: 10.1038/nature14514

Redondo, R., Kim, J., Arons, A., Ramirez, S., Liu, X., and Tonegawa, S. (2014). Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature* 513, 426–430. doi: 10.1038/nature13725

Rescorla, R. (1974). Effect of inflation of the unconditioned stimulus value following conditioning. *J. Comp. Physiol. Psychol.* 86, 101–106. doi: 10.1037/h0035964

Rescorla, R., and Wagener, A. (1972). *A Theory of Pavlovian conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*, New York, NY: Appleton-Century-Crofts.

Richardson, M., Strange, B., and Dolan, R. (2004). Encoding of emotional memories depends on amygdala and hippocampus and their interactions. *Nat. Neurosci.* 7, 278–285. doi: 10.1038/nn1190

Rolls, E., and Grabenhorst, F. (2008). The orbitofrontal cortex and beyond : from affect to decision-making. *Prog. Neurobiol.* 86, 216–244. doi: 10.1016/j.pneurobio.2008.09.001

Roy, M., Piche, M., Chen, J., Peretz, I., and Rainville, P. (2009). Cerebral and spinal modulation of pain by emotions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20900–20905. doi: 10.1073/pnas.0904706106

Sadacca, B. F., Jones, J. L., and Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *Elife* 5:e13665. doi: 10.7554/eLife.13665

Sah, P., Faber, E., Lopez De Armentia, M., and Power, J. (2003). The amygdaloid complex: anatomy and physiology. *Physiol. Rev.* 83, 803–834. doi: 10.1152/physrev.00002.2003

Sangha, S., Chadick, J., and Janak, P. (2013). Safety encoding in the basal amygdala. *J. Neurosci.* 33, 3744–3751. doi: 10.1523/JNEUROSCI.3302-12.2013

Schafer, R., and Oppenheim, A. (2009). *Discrete-Time Signal Processing, 3rd Edn.* Upper Saddle River, NJ: Prentice Hall.

Schoenbaum, G., Chiba, A., and Gallagher, M. (1999). Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *J. Neurosci.* 19, 1876–1884.

Schoenbaum, G., Roesch, M., Stalnaker, T., and Takahashi, Y. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat. Rev. Neurosci.* 10, 885–892. doi: 10.1038/nrn2753

Schultz, D., Balderston, N., Geiger, J., and Helmstetter, F. (2013). Dissociation between implicit and explicit responses in postconditioning UCS revaluation after fear conditioning in humans. *Behav. Neurosci.* 127, 357–368. doi: 10.1037/a0032742

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

Schultz, W. (2000). Multiple reward signals in the brain. *Nat. Rev. Neurosci.* 1, 199–207. doi: 10.1038/35044563

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron* 36, 241–263. doi: 10.1016/S0896-6273(02)00967-4

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.* 57, 87–115. doi: 10.1146/annurev.psych.56.091103.070229

Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Schultz, W., and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* 23, 473–500. doi: 10.1146/annurev.neuro.23.1.473

Scott, D., Stohler, C., Egnatuk, C., Wang, H., Koeppe, R., and Zubieta, J. (2007). Individual differences in reward responding explain placebo-induced expectations and effects. *Neuron* 55, 325–336. doi: 10.1016/j.neuron.2007.06.028

Shabel, S., and Janak, P. (2009). Substantial similarity in amygdala neuronal activity during conditioned appetitive and aversive emotional arousal. *Proc. Natl. Acad. Sci. U.S.A.* 106, 15031–15036. doi: 10.1073/pnas.0905580106

Singer, T. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* 303, 1157–1162. doi: 10.1126/science.1093535

Soltani, A., and Wang, X. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *J. Neurosci.* 26, 3731–3744. doi: 10.1523/JNEUROSCI.5159-05.2006

Soltani, A., and Wang, X. (2010). Synaptic computation underlying probabilistic inference. *Nat. Neurosci.* 13, 112–119. doi: 10.1038/nn.2450

Stalnaker, T., Cooch, N., and Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. *Nat. Neurosci.* 18, 620–627. doi: 10.1038/nn.3982

Steinberg, E., Keiflin, R., Boivin, J., Witten, I., Deisseroth, K., and Janak, P. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* 16, 966–973. doi: 10.1038/nn.3413

Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.* 3, 9–44. doi: 10.1007/BF00115009

Sutton, R., and Barto, A. (1990). "Time-derivative models of pavlovian reinforcement," in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, Vol. 59, eds M. Gabriel and J. Moore (Cambridge: MIT Press), 497–537.

Takahashi, Y., Roesch, M., Stalnaker, T., Haney, R., Calu, D., Taylor, A., et al. (2009). The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. *Neuron* 62, 269–280. doi: 10.1016/j.neuron.2009.03.005

Tovote, P., Fadok, J., and Luthi, A. (2015). Neuronal circuits for fear and anxiety. *Nat. Rev. Neurosci.* 16, 317–331. doi: 10.1038/nrn3945

Uleman, J. S. (1987). Consciousness and control the case of spontaneous trait inferences. *Pers. Soc. Psychol. Bull.* 13, 337–354. doi: 10.1177/0146167287133004

Van Rooij, S., Geuze, E., Kennis, M., Rademaker, A., and Vink, M. (2015). Neural correlates of inhibition and contextual cue processing related to treatment response in PTSD. *Neuropsychopharmacology* 40, 667–675. doi: 10.1038/npp.2014.220

van Vreeswijk, C., and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274, 1724–1726. doi: 10.1126/science.274.5293.1724

van Vreeswijk, C., and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural Comput.* 10, 1321–1371. doi: 10.1162/089976698300017214

Vits, S., Cesko, E., Enck, P., Hillen, U., Schadendorf, D., and Schedlowski, M. (2011). Behavioural conditioning as the mediator of placebo responses in the immune system. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 1799–1807. doi: 10.1098/rstb.2010.0392

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412, 43–48. doi: 10.1038/35083500

Wager, T., Scott, D., and Zubieta, J. (2007). Placebo effects on human mu-opioid activity during pain. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11056–11061. doi: 10.1073/pnas.0702413104

Wagner, G., Koschke, M., Leuf, T., Schlosser, R., and Bar, K. (2009). Reduced heat pain thresholds after sad-mood induction are associated with changes in thalamic activity. *Neuropsychologia* 47, 980–987. doi: 10.1016/j.neuropsychologia.2008.10.021

Wallis, J. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annu. Rev. Neurosci.* 30, 31–56. doi: 10.1146/annurev.neuro.30.051606.094334

Watson, A., El-Deredy, W., Iannetti, G., Lloyd, D., Tracey, I., Vogt, B., et al. (2009). Placebo conditioning and placebo analgesia modulate a common brain network during pain anticipation and perception. *Pain* 145, 24–30. doi: 10.1016/j.pain.2009.04.003

Wiech, K., and Tracey, I. (2009). The influence of negative emotions on pain: behavioral effects and neural mechanisms. *Neuroimage* 47, 987–994. doi: 10.1016/j.neuroimage.2009.05.059

Young, A., Ahier, R., Upton, R., Joseph, M., and Gray, J. (1998). Increased extracellular dopamine in the nucleus accumbens of the rat during associative learning of neutral stimuli. *Neuroscience* 83, 1175–1183. doi: 10.1016/S0306-4522(97)00483-1

Young, R., Cegavske, C., and Thompson, R. (1976). Tone-induced changes in excitability of abducens motoneurons and of the reflex path of nictitating membrane response in rabbit (*Oryctolagus cuniculus*). *J. Comp. Physiol. Psychol.* 90, 424–434. doi: 10.1037/h0077219

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., et al. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science* 323, 1496–1499. doi: 10.1126/science.1167342

Zillmann, D. (1971). Excitation transfer in communication-mediated aggressive behavior. *J. Exp. Soc. Psychol.* 7, 419–434. doi: 10.1016/0022-1031(71)90075-8

Zillmann, D., Katcher, A., and Milavsky, B. (1972). Excitation transfer from physical exercise to subsequent aggressive. *J. Exp. Soc. Psychol.* 8, 247–259. doi: 10.1016/S0022-1031(72)80005-2

Zubieta, J., Bueller, J., Jackson, L., Scott, D., Xu, Y., Koeppe, R., et al. (2005). Placebo effects mediated by endogenous opioid activity on mu-opioid receptors. *J. Neurosci.* 25, 7754–7762. doi: 10.1523/JNEUROSCI.0439-05.2005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.