Mediated Learning: A Computational Rendering of Ketamine-induced Symptoms.

Esther Mondragón

Artificial Intelligence Research Centre (CitAI), Department of Computer Science, City, University of London

Centre for Computational and Animal Learning Research, St. Albans, United Kingdom

Author note. Correspondence should be addressed to Esther Mondragón, Artificial Intelligence Research Centre (CitAI), Department of Computer Science, Northampton Square, City, University of London, London, EC1V 0HB, United Kingdom. E-mail: E.Mondragon@city.ac.uk.

The source code for all simulations can be found at https://github.com/cal-r/DDA_model . Executable versions for different platforms can be downloaded from https://www.cal-r.org/index.php?id=DDA-sim.

Running head: computational ketamine-like effects.

ABSTRACT

This paper explores the contribution of the DDA computational associative learning model to understanding the role of mediated learning mechanisms in the generation of spurious associations, as those postulated to characterize schizophrenia. Three sets of simulations for mediated conditioning, mediated extinction, and a mediated enhancement of latent inhibition, a unique model prediction, are presented. For each set of simulations, a parameter that modulates the impact of associative memory retrieval and the dissipation of non-perceptual activated representations through the network was manipulated. The effect of this operation is analyzed and compared to ketamine-induced effects on associative memories and mediated learning. The model's potential to predict these effects and present a plausible error-correction associative mechanism is discussed in the context of animal models of schizophrenia.

Keywords: associative learning, mediated learning, latent inhibition, computational modeling, associative memories.

Theories of Pavlovian associative learning have typically assumed that associations convey the ability to characterize or represent, in one way or another, the (causal) structure of the environment, learning its regularities (e.g., Mitchell et al., 2009; Rescorla, 1988; Shanks, 2007; Tolman & Brunswik, 1935).

But does the learned structure always match the structure of the world? On the one hand, it is apparent that the connections between different events are not straightforward one-to-one relations; hence, predictions cannot be univocal, resulting in a highly complex mapping between our perception and environmental cues (see Thorn, 2021, for a recent agent simulation and review). On the other hand, learning is not a mere mirroring experience exempt from influences other than external input (Shepard, 1984). I will let aside the problem of the intrinsic complexity of forming an accurate representation of the world to focus on how internal input, in the absence of a corresponding external stimulation, can modulate learning. As we acquire information about the external world's contingencies, existing associations can evoke memories of cues connected to present ones. These retrieved cues can, in turn, enter into or strengthen other associations. The activation of such internal cue representations is often weak, as is the learning they undergo (Hall, 1996).

However, internal input can become prominent in interpreting the world under certain circumstances, generating atypical associations. For instance, it has been suggested that certain neurodevelopment disorders in schizophrenic patients may induce firings and release dopamine independently of stimuli and context. In these individuals, the dopamine system may create abnormal novelty and salience. As a result, delusions can emerge as a top-down explanation to make sense of the peculiar experience. Within this top-down cognitive construction, spurious associations between abnormal salient internal representations and percepts can originate hallucinations (Kapur, 2004). Others (e.g., Gray et al., 1991) highlight an inappropriate prediction error signaling, a product of a disparity

between the expected and actual experiences. Further, incongruous error signaling has been related to disrupted attentional allocation and its direct influence on the associative mechanism (Corlett et al., 2007a).

More directly related to the purpose of this paper and the possible contribution of computational learning models to the study of mental illness is the proposed involvement of mediated learning in the generation of hallucinations (Fleming et al., 2022; Koh & Gallagher, 2020). Mediated learning effects are of interest as they allow for understanding the processes governing learning about memory representations (Holland, 1981, 1983, 1990; Rescorla, 1982). Most learning theoreticians deem mediated learning as a process in which an association, excitatory or inhibitory, is formed between two stimuli that are never directly paired with each other but mediated by the pairing of an associatively activated representation of one of them and the presence of the other (Aitken et al., 2001; Dwyer, 2001; Holland, 1990; Le Pelley & McLaren, 2001; Rescorla, 1982; Ward-Robinson & Hall, 1996). Because such internal input can only be indirectly measured and analyzed, precise formal tools are needed to infer their contribution and assess different postulates. I posit that computational models capable of incorporating into the associative structure retrieved cues and algorithmically describing their interplay and interactions with external stimulation are ideally suited to help us refine our learning theories and drive scientific exploration. Attempts to theoretically account for mediated learning phenomena such as mediated conditioning, mediated extinction, and retrospective revaluation procedures such as backward blocking and unovershadowing (Le Pelley & McLaren, 2001; Miller & Witnauer, 2016) have led researchers to propose ad hoc contradictory learning rules. Wagner (1981) proposed an automatic memory processing model (SOP) instantiated in a real-time elemental connectionist network. Wagner's SOP postulated a compartmental memory system in which stimulus elements are in one of three dynamic states of activation: Inactive

(I), primary activated (A1), or secondary activated (A2). According to this model, stimulus presentation generates elements' A1 activity. As time passes, a proportion of active elements decays passively into the secondary peripheral state A2, and from there, elements decay into an inactive state (I). Crucially, once an association between a predictor and an outcome is established, the former activates the latter's representation into A2. Different rules of learning that define the weight vector direction (positive or excitatory and negative or inhibitory) operate depending on the stimulus role as a predictor or outcome and the states of activation. Thus, in SOP, stimuli in A1 will form mutual excitatory associations; a predictor in A1 will negatively link to an outcome in A2, and, finally, an inhibitory link will be expected to be formed between an outcome stimulus in A1 and predictors in A2. No other rules were explicitly defined by Wagner (1981). Holland (1983) proposed a modification of the original SOP learning rules to allow the formation of an excitatory association between stimuli activated in A2 and in A1. This modification permitted accounting for mediated conditioning. Contrarily, to explain retrospective revaluation effects, Dickinson and Burke (1996) suggested that the association between A2 and A1 should be inhibitory instead and added a further excitatory rule to connect stimuli in A2. A coherent and precise associative learning model of mediated phenomena, capable of generating accurate predictions, is of paramount relevance, particularly for medical research. The Double Error Dynamic Asymptote (DDA) model (Kokkola et al., 2019) posits an error correction learning system in which stimulus elements are activated and their weights updated in a time continuum independently of whether the input is external or internal (retrieved by an existing association). Unlike any other model, the asymptote of learning against which the prediction error is evaluated varies with the activity of the connected stimuli's nodes, according to a similarity function between the cues' activities. Thus, the model encapsulates the biological principles of the Hebbian rule while keeping a

fundamental associative error correction mechanism. As a result, the direction of learning between two cues is determined solely by the discrepancy between the maximal learning achievable for a pair of cues and the total prediction of the outcome stimulus at any given time. This principle endows the model with the capability of accounting for mediated learning and retrospective revaluations with a single and consistent learning rule (for a review of the DDA model's simulation scope, see Kokkola et al., 2019).

Furthermore, attention and learning, and, critically, their mutual interaction, have become a prominent area of study in the clinical context (Baruch et al., 1988; Lubow & Weiner, 2010). Associative learning theories of attention postulate a direct relationship between a stimulus's effective salience and selective attention. Since their origin, selective attention models rely on the idea of limited computational resources to process the inputs and are conceptualized as a bottleneck tactic response to these constraints that filters out all but the relevant stimuli to solve a task (Broadbent, 1958; Treisman, 1964). From the perspective of associative learning, the effective salience of a cue, the main feature that drives attention, modulates, and it is, in turn, modulated by the predictive value of the stimulus. Traditional attentional learning models, however, only consider directly activated stimuli when describing attention modulation of learning (Le Pelley et al., 2016, for a review). Some discrete associative learning models assume that the effective salience of a stimulus to enter into an association (Esber & Haselgrove, 2011; Pearce & Hall, 1980) or its expression (Miller & Matzel, 1988) decays with exposure. These models adopt binary activation of stimuli (present or absent), and exposure is, in practice, identified to the number of direct activations (events or trials) (for a review, Schmajuk & Alonso, 2012).

Conversely, real- or discretized-time computational models (e.g., Kokkola et al., 2019; Kutlu & Schmajuk, 2012; Mondragón et al., 2013; Mondragón et al., 2014; Schmajuk et al., 1996; Sutton & Barto, 1987) postulate a graded pattern of stimulus activation. Such representation enables updating learning and the stimulus' effective salience at each time step. The update occurs during the stimulus presentation and after its offset, while the activation decay pattern remains above a certain, more or less arbitrary, threshold. Other models go further and implement a connectionist network that incorporates learning between a present cue and non-directly activated but associatively retrieved stimuli (e.g., Brandon et al., 2000; Kokkola et al., 2019; Wagner, 1981; Wagner & Brandon, 2001). Such computational models may significantly broaden our understanding of the role of internal memory representations in generating aberrant behaviors in mental illness. In particular, in the DDA model, attention modulation is updated at each time step, changing with the proximity to the outcome. Crucially, the model postulates attentional variation to both active and associatively retrieved predictors. Hence, the DDA model sustains that attention to a present cue will also be affected by the number of cue-retrieved but absent elements –which will, in turn, undergo changes in their effective salience during associative activation.

Associative learning has a long-standing influence on clinical studies and treatment and is still one of the main paradigms for understanding the basic principles of human pathological behavior (Corlett & Schoenbaum, 2021; Haselgrove & Hogarth, 2011; Lewis et al., 2013; Schachtman & Reilly, 2011). In this paper, I present simulations of the DDA model (Kokkola et al., 2019) that aim to replicate some empirical manipulations that produce psychosis-like effects in rodents. The DDA model, unlike other theories that encompass algorithms for associatively activated cue learning (e.g., Brandon et al., 2000; Wagner, 1981; Wagner & Brandon, 2001; but see McLaren et al., 1989; McLaren, 2011)

incorporates computational mechanisms to integrate attention and learning for memory retrieved representations. In addition, to my knowledge, only the DDA model has succeeded in explaining mediated effects with a single associative learning rule instead of postulating different, potentially conflicting rules for different phenomena.

METHODS

Transparency and Openness

No data was used in this paper. Additional details and formalisms of the DDA model were published in Kokkola, Mondragón, and Alonso (2019). Executable files and simulators of the model for different platforms (Kokkola, Mondragón, and Alonso (2018) are available for download at https://www.cal-r.org/index.php?id=DDA-sim. The code is also available at https://github.com/cal-r/DDA_model. The simulations and designs in this study were not preregistered.

The DDA Model

DDA is a fully connected network architecture instantiated as a discretized-time computational model of associative learning. The model posits multiple nodes that represent stimulus elements, which can be unique to a stimulus or shared. Elements are clustered in time. Individual element activity is binary (active or not), but the cluster's activity is given by the mean number of its active elements. Critically, clusters of elements are not only directly activated through external sensory input but also associatively, retrieved by their association with other elements (internal activation). If a cluster is both directly and associatively activated, its overall activation is assumed to be whichever is larger. When two or more clusters are activated, independently of the source, the elements of one node enter into association with other active clusters. The model's learning algorithm is error

correction based, and, as in the case of Rescorla-Wagner (Rescorla & Wagner, 1972), it takes into account the total prediction of a given outcome. Two main features differentiate the DDA learning algorithm from other global error correction models. First, inspired by the Hebbian assumption (Hebb, 1949), the asymptote of learning $\lambda_{i\rightarrow j}^{t}$ does not adopt a fixed value depending on the binary activation of the outcome (1 or 0) but a variable one instead. This value is estimated as a function of the non-binary activity of the two clusters involved, the predictor and the outcome. More specifically, at each time step, the asymptote of learning is calculated as an inverse measure of the distance in activity between the predictor cluster $\hat{A}_i^t$ and the predicted cluster of elements $\hat{A}_j^t$. That is, the larger the discrepancy, the lower the asymptote. The exact computation, however, renders the contribution of the predictor and outcome activities anti-symmetrical, causing the outcome activity to be more determinant of whether the asymptote is positive or negative. Formally,

$$\lambda_{i\rightarrow j}^{t} \triangleq \frac{\hat{A}_j^t - \left|\hat{A}_j^t - \hat{A}_i^t\right|}{\max\left(\hat{A}_j^t, \hat{A}_i^t\right)}, \tag{1}$$

where $\hat{A}_i^t$ represents the overall activation of stimulus *i*.

Second, the weights update is not only a function of the prediction error (the error in predicting the outcome, $\delta_{i\rightarrow j}^{t}$) but also a function of the predictor's error (the error in anticipating the predictor, $\delta_{\rightarrow i}^{t}$). The latter captures the notion that learning is modulated by the familiarity of cues in a context, incorporating a perception of irregularity or instability to the predictive element incidence. The more familiar a cue is, the less urgent it is to learn about its consequences. This postulate delays the predictor-outcome association –an analogous effect to that theorized by Wagner's SOP A2 activation state (Wagner, 1981). However, unlike Wagner's conceptualization, it does not hinder the element activation or its

effective processing. Instead, it facilitates the formation of the predictor's representation by activating a greater number of its elements, fostering elements' associations in a process akin to unitization (McLaren & Mackintosh, 2000). Yet, in contrast with the assumption of the McLaren and Mackintosh model, the predictor's error operates by modulating the prediction error, not by boosting the node activation, which remains unchanged –hence the appellative of 'Double Error' (D), defined as

$$\mathrm{D}^{t}_{i\to j} \simeq \delta^{t}_{\to i} \cdot \delta^{t}_{i\to j} \cdot x^{t}_{i\vee j} \ , \tag{1}$$

where $x^{t}_{i\vee j}$ represents the binary activation of elements *i* or *j*.

As an attentional learning model, the DDA model borrows Pearce-Hall's notion that attention to cues quantified as a learning rate parameter called associability (α)increases with the uncertainty of the occurrence of an outcome (Pearce & Hall, 1980). It differs from Pearce-Hall's in that associability in the DDA model is updated in a real-time framework. More importantly, this attentional process operates independently of whether the outcome is a motivationally relevant or a neutral stimulus and whether the cue is externally activated or associatively retrieved. The model further assumes that if the uncertainty remains high over a sufficient period, the agent uses this persistent uncertainty as a source of regular information, reducing their attention to the cue.

Learning is, thus, driven by the double error term, the cues' associability at each timestep ($\alpha^{t}_{i,}$), the physical intensity of the stimulus involved (*s*), and their current level of activation (A^{t}).

$$\Delta w_{i\to j}^{t} \simeq s_i \cdot s_j \cdot \ \alpha_{i,}^{t} \cdot \mathrm{D}_{i\to j}^{t}\, A_i^t \cdot A_j^t \qquad (2)$$

Not only do associatively activated nodes undergo learning, but more significantly, they are also subject to the same changes in attention (or associability) that directly activated nodes undertake. Variations in attentional modulation are regulated by α and are proportional to the stimulus activation, the prediction error (Equation 4), and a parameter ρ that determines the speed of change. Algorithmically, the model distinguishes between two attentional parameters depending on whether the outcome is an unconditioned stimulus (US) or a neutral stimulus (α_r and α_n, respectively). Similar to Pearce and Hall's (1980) α, uncertainty in the occurrence of an outcome accelerates learning toward it. However, α in the DDA model differs from Pearce and Hall's (1980) in several ways. First, updates are computed each time step and for all active stimuli, regardless of whether they are externally or associatively activated. Second, if this uncertainty remains for an extended period, the level of attention decays again. Third, it is calculated for all outcomes. Lastly, for each outcome class, a cue's associability changes influence its attention toward other outcomes.

Since any element in the network can induce activity in an associated cluster, propagation through the network can be activated directly from the activity engendered by external sensory stimulation and associatively by an active internal element associated with it. The propagation of internal activation is, nonetheless, attenuated by dissipation. A parameter theta (ϑ) regulates the degree of dissipation of non-externally activated clusters, critically conferring a greater contribution to direct element input. In other words, ϑ modulates the spread of the retrieved activation generated by direct or by other associatively retrieved cues, multiplicatively dissipating the activation deeper in the network –effectively controlling stimulus activation without external stimulation. To avoid infinite reverberation of activity through the network, as discussed in (Wagner, 1981, p. 13), ϑ should adopt a

value smaller than 1, thus reducing the activation of non-directly input activation. That is, whenever the predictor cluster is not present but associatively retrieved, the outcome prediction Ψ_j^t within the error prediction term is modulated by ϑ.

$$\delta_{i\to j}^t = \lambda_{i\to j}^t - \Psi_j^t \cdot (\vartheta \vee 1) \tag{3}$$

The model's predicted response is calculated as the aggregated US predictions at each timestep since all contribute to the associative activation of the US clusters. For a given trial (T), the predicted response is calculated by averaging the above values over time and normalized between 0 and 100.

$$R_{(\to \mathrm{US})}^T \simeq \sum_{i(t\in T)\in \vec{+}} \frac{\Psi_{i\to +}^t}{\#t \in T} \tag{4}$$

The reader is referred to the original paper for a more detailed description of the model (Kokkola et al., 2019).

Simulations Assumptions and General Parameters

It has been reported in the literature (e.g., Fleming et al., 2022; Koh et al., 2018) that administrations of ketamine, a substance that intervenes in the regulation of dopamine (Marcus & Bruchas, 2021), can produce hallucinations-like symptoms in rodents and increase mediated conditioning effects. I hypothesize that atypical mediated effects that follow the administration of ketamine can be simulated by increasing the values of the DDA model ϑ hyperparameter, which intensifies the reverberation of associative activation in the network. As discussed earlier, normal learning requires $\vartheta < 1$. The model described in Kokkola et al. (2019) adopted $\vartheta = 0.95$ as a standard value to successfully simulate an

extensive range of phenomena. Thus, the model's ability to predict and alter the strength of mediated learning was explored under different ϑ conditions (or groups), namely, Standard ($\vartheta = 0.95$), Raised ($\vartheta = 0.975$), and High ($\vartheta = 1.05$). Group Raised defines an increment over the standard value that is nonetheless below the threshold. Group High describes a value that is above the threshold. The initial assumption was that moderated increases in theta would favor mediated learning. A larger increase may result in infinite reverberation, disrupting normal learning. In the next section, simulations of mediated conditioning, mediated extinction, and a DDA model prediction of mediated latent inhibition will be presented, and the learning interactions between associatively activated cues and their attentional fluctuations will be analyzed.

Simulations were carried out with the DDA Simulator Ver.1, publicly available at https://www.cal-r.org/index.php?id=DDA-sim. The code is deposited in GitHub, https://github.com/cal-r/DDA_model. The model fixed hyperparameters used are given in Table 1.

TABLE 1 ABOUT HERE

All stimuli in these simulations had a duration of 3 time-units (t), except for the US, which had a length of 1t, and presentations were distributed separated by a 100t inter-trial interval. To control order bias, whenever two or more stimuli were presented within a phase, the simulation ran a minimum of 30 different random sequences and averaged the values per stimulus and trial. The US salience 1, context salience 0.001, stimulus and context initial associability towards the US and neutral stimuli ($\alpha_r = 1$ and $\alpha_n = 0.8$, context_$\alpha_r = \alpha_n = 0.01$) remained unchanged across the simulations. The stimulus salience and the associability recency (ρ), which regulates the amount of attention change across trials, were set to 0.25 and 0.01, respectively, in Experiments 1 and 2. The stimulus salience was reduced to 0.075

and ρ set to 0.1 in Experiment 3 to slow down learning and facilitate observing a sizable latent inhibition effect, t.

To parallel a relationship between response measures that are inversely proportional to the predictions (i.e., the higher the association, the lower the response), a ratio similar to the suppression ratio used in many empirical settings was computed (based on Mondragón et al., 2014):

$$r = \frac{\left(\max R^T - R_i^T\right)^2}{\left(\max R^T - R_i^T\right)^2 + \left(\max R^T - R_j^T\right)^2}, \tag{6}$$

where R_i^T is the simulated conditioned response to cue *i* (the weakest predictor) at trial *T*; $\max R^T$ the maximum response value (assumed to be 100) and R_j^T is the simulated conditioned response to cue *j* (the strongest predictor) at trial *T*.

SIMULATIONS AND RESULTS

In this section, I present a set of simulations aimed at reproducing some mediated phenomena and a unique model prediction to explore the effect of varying the magnitude of the retrieval of associatively activated cues. This novel prediction is included to illustrate how mediated effects could potentially interact with the stimulus's associability when attention is assumed to be the decisive contributor to the expected result. In all the simulations, the model's ϑ hyperparameter was manipulated, and the effect on the weight of the mediated associations and tests predicted responses were assessed. Table 2 shows the design used in the simulation of each phenomenon.

TABLE 2 ABOUT HERE

Simulation Set 1: Mediated Conditioning

A within-subjects design was used to simulate a mediated conditioning experiment in three independent conditions (Group Standard, Group Raised, and Group High) that differed only on the value of the parameter ϑ with which the simulation was run. Phase 1 consisted of 40 trials. Two simultaneous compound stimuli, namely AB and XY, were presented each 20 times in a semi-random manner in Phase 1. During Phase 2, one stimulus of each compound, B and Y, was introduced 16 times. Stimulus B was consistently paired with a US, whereas stimulus Y was presented alone. Phase 3, a Mediated Conditioning (MC) test, consisted of a single presentation of stimulus A and of stimulus X. Finally, a Direct Conditioning test was given to assess the learning acquired by stimuli B and Y.

This design can be considered a conceptual replication of Fleming et al. (2022), which trained rats in a mediated learning avoidance task. In their experiment, two combinations of flavor-odor stimuli were presented in Phase 1. In Phase 2, one of the flavors was evaluated by conditioning it with LiCl, whereas the other was paired with saline. An injection of ketamine preceded the devaluated flavor but not the one paired with saline. The dose of ketamine varied for the different groups: 0mg/kg, 10mg/kg, or 30 mg/kg. A test was then carried out in which differential consumption of odored water was measured. Their design purposely administered ketamine selectively to the conditioned stimulus during the devaluation phase to dissociate potential effects produced by the formation of perceptual expectations from their retrieved representation.

Unlike in Fleming et al.'s (2022) experiment, the ϑ manipulation in this simulation applied across phases. A reason for this choice is that ϑ is a hyperparameter of the model

designed to operate throughout the network training. Changing it within phases and conditions would be an ad hoc model modification. Another is that, although the manipulation does not match the procedure required to empirically dissociate the formation of perceptual expectations from changes in the stimulus retrieved representation as intended by Fleming et al. (2022), the distinction is conceptually irrelevant from the model's point of view. Any effect of the manipulation of ϑ within the model excludes an interpretation in terms of the formation or retrieval of perceptual expectations since, unlike in the case of drug manipulation, no outcoming perceptual effects are generated –other than the resonation of the stimuli's elements. The parameter ϑ simply regulates the retrieval of an association and its spread inside the network, intensifying the internal reverberation of associatively activated elements as its value increases. As a result, at each trial, the extent and depth of the associative activation are boosted, favoring mediated conditioning.

FIGURE 1 ABOUT HERE

To compare and evaluate the results of this conceptual replication with those reported by Fleming et al. (2022), we computed an avoidance ratio with the simulated data. The right panel of Figure 1 shows the simulated results alongside an adaptation of Fleming et al. (2022) results displayed in the left panel. The three columns on the left hand-side of each panel depict the computed avoidance ratio during the MC test for the three independent treatment conditions (Ketamine dose in the left panel and ϑ value in the right panel). The three columns on the right-hand side of each panel show the corresponding paired stimuli during the Direct Conditioning test. It is apparent that the model results match the pattern of the empirical results. However, to account for it, the DDA model does not require to appeal to perceptual expectations. It offers a good conceptual simulation for an enhancement of mediated conditioning and an alternative analysis to the observed empirical results.

Similar to the effect observed under an intermediate ketamine dose in Fleming et al. (2022) experiment, a raised ϑ value ($\vartheta = 0.975$) enhanced mediated conditioning in comparison to that empirically observed with a sham dose of ketamine and simulated with the standard ϑ value ($\vartheta = 0.95$), respectively. When a higher than 1 value of ϑ was employed, direct and, consequently, mediated conditioning were impaired, reproducing the empirical data observed with the largest dose of ketamine.

The effect of manipulating ϑ on the magnitude of mediated conditioning is evident when looking at the simulated conditioned response for the mediated conditioned test stimulus A and the control stimulus X, shown in the right panel of Figure 2. In all three treatment conditions, the response to A is predicted to be higher than to the control stimulus X, for which no response was predicted. In addition, the simulation predicted a more robust conditioned response to A in Group Raised than in Group Standard and Group High, with the latter generating the weakest response. Per the model specification, the simulated response results from aggregating all US predictions. We need to look at the corresponding weights to determine the specific contribution of a given cue-US association. The left panel of Figure 2 displays the simulated weights of A and X links to the US at each training trial. The pattern of results is equivalent to that described for the conditioned response, suggesting that cue-US associations are the main contributors to the simulated response. It is worth noticing that the negative value of the X to US weight in Group High indicates the development of weak inhibitory learning between the putative neutral cue and the US.

FIGURE 2 ABOUT HERE

A simulation allows us to explore further the source of the effect, which will be hard to address behaviorally. The top panel of Figure 3 displays the learning weights of the

directly conditioned B and the control Y to the US and the mediated conditioning acquired by A and the control X during Phase 2 per group. A quick inspection reveals that the rate of direct conditioning to B decreases as ϑ increases from 0.95 to 1.05. However, a stronger mediated conditioning to A was observed in Group Raised compared to Group Standard and Group High, for which the weights considerably decreased as training progressed. The learning rate parameter α, which quantifies the cue's associability and is assumed to reflect the level of attention, was assessed next. The α values for cue-US links toward a motivational stimulus are presented in the middle panel of Figure 3. Stimulus B, which was directly activated and underwent conditioning to the US, exhibited the highest α values. The associability of A, the mediated conditioned stimulus, held an intermediate value. Finally, the non-rewarded stimuli, X and Y, had the lowest associability level. This pattern was constant throughout trials, with little variation across groups.

FIGURE 3 ABOUT HERE

Mediated conditioning strongly depends on the degree of retrieval of the neutral cue (A) to its retriever (B) at the time of conditioning. The bottom panel of Figure 3 shows the connection weights from the directedly activated cues (B and Y) to their linked stimuli (A and X). The initial associative weights between the neutral stimuli are moderately higher in Group Raised than in the other groups. More critically, as the value of ϑ increased, the rate of extinction of these associations slowed down. This effect was particularly evident for the B-A link (white circles). It is worth noticing, too, that resistance to extinction was more marked when the stimuli involved were not directly or associatively paired to the US (black cross).

As the associative activation of clusters of elements by other retrieved elements is intensified by the increase in the ϑ parameter, more stimulus elements overpass the threshold activation and are recruited into activation. More activation implies stronger conditioning. However, associative retrieval is a function of the connecting weights, which are expected to undergo some extinction, lowering the activation level of the retrieved elements compared to the activation of directly input stimuli. This discrepancy in activity limits the amount of learning in the DDA model, which posits a dynamic asymptote that is inversely proportional to the distance in activation between the involved elements. Under these conditions, the robust initial association between B and A and Y and X is expected to be relatively sustained due to the reverberation produced by similarly strongly activated retrieved elements. In US trials, an elevated ϑ also prompts that more US elements are recruited into associative activation, thus competing, due to bidirectional associations, as predictors of the neutral outcome. Cue competition reduces the magnitude of the protection from extinction between the neutral cues. Although lessened, the association between B and A persists, and it is stronger with raised ϑ values, guaranteeing that more elements of A are recruited during conditioning, increasing the chances of acquiring an association with the US. Much larger ϑ values, though, boost associative retrieval reverberation to a point at which too many elements are simultaneously active, including elements of the directly activated stimuli, both predictor and US. As a consequence, multiple associations between retrieved elements and directly activated elements take place with a lower asymptotic value, which results in low or negative predictions. These predictions contribute to the total cue prediction, reducing the total amount of conditioning –a form of cue competition between directly active stimulus' and retrieved stimulus elements that can potentially disrupt direct and mediated conditioning alike.

In summary, a simple error correction mechanism that incorporates all active (not just present) predictors and a boost in activation retrieval by associatively active elements as formalized in the DDA model can account for the mediated conditioning changes reported by Fleming et al. (2022). Although a critical difference exists between the empirical setting and the reported simulations (in the former, ketamine was administered only prior to the devaluation; in the simulations, the parameter change was effective throughout all phases), the procedural difference does not bear significance from the model's perspective. The manipulation in the model excludes the formation of perceptual expectations and capitalizes on associative retrieval during mediated conditioning, effectively intensifying the reverberation of absent but cue-retrieved elements. Raised values of ϑ enhance the span of associative activation of retrieved elements, favoring the opportunities for mediated conditioning. High values disrupt normal learning in a mediated conditioning setting, possibly inducing infinity reverberation of the retrieved elements' activation in the network, as discussed in Wagner (1981).

Attentional factors, as captured by the associability parameter α, do not seem to impact the observed mediated conditioning enhancement.

Simulation Set 2: Mediated Extinction

The following simulations used a within-subjects design to replicate a generic mediated extinction effect. The design used is comparable to the one employed in the previous simulation set to keep consistency and facilitate analysis. Three independent groups (Group Standard, Group Raised, and Group High) were run. The experiment consisted of four phases. In Phase 1, 20 trials of a simultaneous compound AB and 20 trials of XY were presented semi-randomly. In Phase 2, one stimulus of each compound, A and X, was each paired with the US 20 times. During Phase 3, Mediated Extinction, stimulus B

was presented in isolation in 20 trials. Finally, Phase 4 consisted of a single test presentation of stimulus A and of stimulus X.

Figure 4 shows the test results. The simulated response (right panel) to the mediated extinguished stimulus A was less pronounced than to the control stimulus X in all three groups. The response to A decreased as ϑ increased, but despite this reduction, the difference between the cues, which indicates mediated extinction, seems to be largest in Group Raised and smallest in Group High –which may indicate an attenuation of extinction with large ϑ values. The same pattern is observed when looking at the cue-US weights (left panel) across all trials, which replicates the direction of the magnitude of mediated learning previously simulated for mediated conditioning. The analysis is, however, more complex than in the previous case. Assuming equal learning mechanisms between the two simulations, A-US and X-US pairings during Phase 2 are expected to generate mediated conditioning to their corresponding associated cues, B and Y. Subsequent extinction of B should, therefore, be regulated by the previously acquired mediated conditioning of B. Thus, the magnitude of the mediated extinction to A should, in turn, depend on the extent to which B is capable of retrieving the US representation. The top panel of Figure 5 shows the weights of the directly conditioned A-US and X-US associations and the magnitude of mediated conditioning acquired by their respectively paired cues, B and Y. This pattern of results fully replicates those simulated previously, showing that the rate of direct conditioning decreases as ϑ increases. Mediated conditioning was also most robust in Group Raised and weakest in Group High.

FIGURE 4 ABOUT HERE

The associative weights of the links connecting the stimuli to the US in Phase 3, during which mediated extinction took place, are displayed in the middle panel of Figure 5. Initial conditioning levels differed at each condition for both A, which was directly conditioned in the previous phase and underwent mediated extinction in Phase 3, and B, which acquired mediated conditioning in Phase 2 and was extinguished during Phase 3. Despite this difference, a distinctive rate of extinction emerged. Stimulus B, which was never directly paired to the US, extinguished at a similar pace in Groups Standard and Raised. However, the level of conditioning of B at the start of the extinction training was lower in the former, and weights became negative early during training compared to Group Raised, in which they remained positive. The associative weights in Group High, which held the lowest conditioning level, quickly approached zero. More significant is the pattern of mediated extinction of A, which was faster and more profound in Group Raised than in the others. The rate of mediated extinction in Group High was moderately slower than in Group Standard, but the final amount of extinction was, if anything, slightly more pronounced.

FIGURE 5 ABOUT HERE

The bottom panel of Figure 5 displays the weights between the directly extinguished cue B and the retrieved stimulus A across the mediated extinction training per group. Extinction of the neutral cues' association was less effective as ϑ increased. High ϑ values resulted in more A and B elements being recruited into activation by associative retrieval (by self-activation or by a liked stimulus). This reverberation of the activation through the network conveys concurrent activation of elements with a similar activity, which translates into effective associations, protecting the neutral cues' association from mutual extinction.

As a result, a high number of A elements are retrieved during B extinction trials, which is expected to intensify the extinction of the A-US link. However, potential cue competition between direct and retrieved elements with high ϑ, as argued in the previous simulation set, is less significant in this scenario where only one stimulus was directly activated. Hence, with more active elements, the B-A association was sustained in Group High, remaining stronger than in the other groups. Yet, because the initial B-US mediated conditioning in Group High was the lowest, the available associative space for mediated US extinction was limited in this group.

In summary, mediated extinction simulations suggest that intermediate values of ϑ engender a faster and more profound mediated extinction effect within the parameter range employed. On the contrary, larger values of ϑ seem to attenuate extinction. However, the latter could be due to an initial disruption of mediated conditioning. Boosted element associative retrieval and simple error correction mechanisms can account for enhanced mediated learning. Koh et al. (2018) found a reduction of extinction in mice treated with ketamine chronically administered prior to training. Their experiment, however, used a discrimination training instead of simple mediated conditioning, which exposes animals to the control cue in the presence of associative memories of the US retrieved by the context during mediated training. Likewise, the associative retrieval of the experimental cue takes place in the absence of the US. It is thus conceivable that this discrimination training may result in a partial reinforcement extinction effect, which is expected to delay extinction.

Simulation Set 3: Mediated Latent Inhibition

Latent inhibition refers to a delay in learning about a familiar cue. In a lab-controlled situation, latent inhibition involves preexposing a target cue before conditioning and comparing this cue's rate of acquisition with the conditioning rate to a non-preexposed stimulus. Attentional models of learning assume that latent inhibition is mostly driven by a loss in a stimulus' effective salience during preexposure (Kokkola et al., 2019; Pearce & Hall, 1980; McLaren & Mackintosh, 2000; Schmajuk et al., 1996) and it is believed to be an adaptive mechanism to ignore inconsequential stimuli (Kaplan & Lubow, 2011). Latent inhibition has been extensively researched in the context of clinical studies of schizophrenia (Lubow & Weiner, 2010; Rascle et al., 2001) and has recently been reassessed as a possible biomarker (Granger et al., 2020). Although other theoretical approaches have been proposed in the literature (e.g., Bouton, 1993; Miller et al., 1986; Wagner, 1981), most accounts of latent inhibition situate attention mechanisms at the center of their analysis.

The design presented here aims to test a potential mediated enhancement of latent inhibition to a cue by associative retrieval of a paired stimulus. According to the DDA model, the reverberation of retrieved cues in a network effectively increases the opportunity for learning –thus enhancing mediated conditioning and mediated extinction (Simulation sets 1 and 2). For attentional models, the opportunity for learning conveys changes in the stimulus associability. If so, attention to cues in memory that are susceptible to undergoing changes in their associative value should also vary. However, unlike the DDA model, other attentional models only consider attention to physically present cues. The rationale to include this model prediction precisely derives from the model conceptualization of learning and attention modulation operating for all active cues, disregarding whether they are directly or associatively activated. We aim to present a testable prediction for a plausible interaction between mediated learning and the associability of absent cues. In Phase 1, two pairs of compound stimuli, AB and XY, were preexposed in a semi-random manner, 10 times each,

as in previous simulations. During Phase 2, 50 non-reinforced presentations of stimulus B were given. Phase 3 consisted of 20 conditioning trials to A, X, and a novel stimulus C. Conditioning to X was expected to be slower than to the novel stimulus C (a non-preexposed control). In addition, if Phase 2 was effective in mediately enhancing latent inhibition to A, conditioning to the latter should be further delayed in comparison to X. Three independent runs of the same design were conducted varying the ϑ parameter in groups Standard, Raised, and High.

FIGURE 6 ABOUT HERE

The simulated conditioned response during Phase 3 to stimuli A, X, and C is displayed in the bottom panel of Figure 6. As expected, conditioning to the novel stimulus C developed much more readily than to the other cues in all groups, with a slightly higher asymptotic level in Group High. Conditioning to cue X that underwent preexposure training in compound was, in turn, faster than to A. The difference in the rate of learning between X and C simulates a latent inhibition effect following compound preexposure (Honey & Hall, 1989; Leung et al., 2011). The fact that conditioning to X was faster than to A requires further analysis. Compound training during Phase 1 is expected to establish an association between the compound stimuli. As a result, in Phase 2, stimulus A is associatively retrieved during B trials. Under these conditions, as hypothesized, A will receive mediated extra exposure, ensuring a more substantial delay in conditioning than that observed for X, which was not preexposed further. I have labeled this enhancement in the amount of latent inhibition as mediated latent inhibition (MLI). To my knowledge, this is a DDA model prediction that has not been reported in the literature. MLI was replicated in each group, with relatively few differences, although it appears to be stronger in Group Standard. The

weights of the A-US, X-US and C-US links across all 60 trials (Figure 6, top panel) followed a similar graded pattern, with the mediated cue A displaying the slowest rate of conditioning, followed by X that, in turn, conditioned at a lower speed than the novel cue C. No clear difference was noticeable between the groups in the rate of conditioning to A or in the rate of extinction of the retriever-retrieved stimulus link (Figure 7 left panel). However, latent inhibition to X seems to increase with ϑ being more profound in Group Raised. A ratio of the X-US over C-US weights was calculated and shown in Figure 7 right panel to highlight the differences in the predictive value of X at each group. This ratio quantifies the strength of the latent inhibition effect. It is apparent that early during training, conditioning to X in Group Raised was substantially more delayed than in groups Standard and High, and the latter proceeded more slowly than in Group Standard.

FIGURE 7 ABOUT HERE

We need to look inside the network to find the source of these differences. The parameter ϑ modulates the spread of associative activation in the network. We have seen that it induces changes in the strength of the stimulus associations by recruiting more elements into activity and making them available for conditioning. However, any critical effect in generating latent inhibition must be the result of the differential preexposure treatment in which no US is present. There are at least two other processes that can be affected by changes in ϑ: the associations between elements and attentional changes. The top left panel of Figure 8 shows the values of the attentional parameter, α, per cue and group at the end of Phase 1. Compound preexposure resulted in a reduction of α from an initial value of 1 to near 0.3. There were no remarkable differences between cues and groups other

than those derived from the randomization of trials. The center and right panels of Figure 8 display the changes of α across groups for cues A and B, respectively. The effective salience of B, which was directly exposed, decayed progressively with trials, and this decrease was equivalent in all groups. The effective salience of the mediated activated cue A was also reduced, and this decrement was slightly more pronounced in Group High. The bottom panel of Figure 8 shows the variations in α during Phase 3 per group and trial. Cues A, X, and C are presented in individual panels, from left to right. The associability of cue A, which underwent mediated preexposure, was the lowest of all cues at the beginning of training. It increased early during the acquisition, reaching an intermediate plateau value afterward. The amount of growth followed, to some extent, a decreasing gradation in relation to ϑ. Thus, α increased at a slower pace in Group High than in Group Raised, and, in the latter, the increment in α was marginally delayed compared to Group Standard. Stimulus X, which received compound preexposure (center panel), had a higher initial associability than A. There were no substantial changes in the effective salience of X during training and no obvious differences between groups. The associability of the novel stimulus C (right panel) decayed from an initial value of 1 across trials. The rate of decay was substantially deeper in Group High than in the other two groups. As more elements are recruited into associative activation, conditioning is expected to strengthen (as observed in the simulation), increasing the predictiveness of the cue, which should result in a loss of associability.

FIGURE 8 ABOUT HERE

The variations in the associability of the stimuli in these simulations, assumed to reflect attentional mechanisms, can solely account for the different rates of US conditioning in Phase 1. First, attention to X and A was reduced during compound preexposure. During the mediated exposure phase, the effective salience of the non-directly presented A was further reduced by associative activation by B. As a result, attention to these cues, as measured by α, diverged at the start of conditioning, with lower α for the mediated preexposed stimulus A, reducing the rate of acquisition of this cue.

It remains to explain why latent inhibition to X is increased in Group Raised. Variations in attention do not seem to account for this difference. Tracing the source of this effect requires looking at the associations formed during preexposure. As ϑ increases, so does the proportion of elements associatively activated, and a greater number of associations between elements of the same stimulus can be formed, which could account for a more robust latent inhibition effect. At the end of Phase 1, no obvious differences in the weights of the X-X stimulus association were found among groups (0.002753, 0.00288, 0.00276, per Group Standard, Group Raised, and Group High, respectively). It must be taken into account that this computation averages the elements' weights to determine their contribution to the stimulus prediction; it does not tell us the number of active elements. We have argued that intermediate ϑ values were likely to result in a substantial number of associatively activated elements becoming available for conditioning. However, according to the DDA model, the high discrepancy in activity between the associatively activated elements and the directly activated US will yield a low asymptotic value. Error correction between a low asymptote and a large total US prediction would render the weaker association between the retrieved elements and the US inhibitory. Since these elements contribute to the total US prediction, they will reduce the US predictive value at each learning episode, thus delaying conditioning. As in the previous simulated scenarios, more significant increases in ϑ may

result in cue competition between associatively activated and directly activated elements, reducing the contribution of mediated learning toward the direct outcome.

It must be noticed that, unlike in previous simulations, ϑ did not affect the rate of extinction between cues. One possible reason is the initial lower association resulting from compound preexposure, which was shorter. In addition, a much higher recency modulator for attention (ρ) and less salient stimuli were used to observe a broader contribution of trials and a more gradual attentional change.

In summary, the simulation produced a robust latent inhibition effect to a cue preexposed in compound relative to a novel stimulus. The magnitude of this effect was enhanced when the target cue was further exposed mediately by preexposing its associated cue. MLI was not substantially affected by manipulating the parameter ϑ, but latent inhibition was somehow more substantial in Group Raised. One possible reason for not observing an effect of ϑ could be due to the fact that MLI takes place after the stimulus has already lost associability and for an extended period –which results in substantial extinction of the neutral stimuli association. Consequently, retrieval of the mediated stimulus is weakened, and the boost produced by increasing theta is less effective. Another more theoretically relevant one is that LI is assumed to critically depend on the attentional factors –which, from the point of view of the attentional models of learning, are only an indirect effect of the associative process.

DISCUSSION

The paper reports simulations of mediated conditioning, mediated extinction, and a mediated enhancement effect of latent inhibition. For each phenomenon, three independent simulations were performed to analyze the effect of varying a DDA model parameter, ϑ, that regulates the amount of associative retrieval by non-directly activated cues. The parameter modulates the spread of activation throughout the network. Increasing its value effectively boosts the number of active elements of a given mediated representation. In other words, ϑ regulates the degree to which memories of non-present events reverberate in the network. A larger number of active retrieved elements is expected to increase mediated learning, which can result in excitatory or inhibitory connections depending on the discrepancy in the level of activation of the cluster of elements –upon which the current maximal outcome value is computed– and the extant total available prediction. Individual mediated element associations summate to the total outcome prediction. Larger ϑ values promote competition between retrieved and direct input elements. The combination of these mechanisms is determinant of the predicted response.

Consistent with the above, mediated conditioning was strongest when a moderately raised ϑ was used and weakest when further increased. Mediated extinction was also more pronounced when a moderately increased ϑ value was used in the simulations. Despite the necessary design differences, both sets of simulations replicated the enhanced mediated conditioning pattern. Comparable results were also observed for mediated extinction, which was achieved faster and more profoundly with a modest ϑ increment. These results are consistent with the rationale above. As ϑ increased, so did the protection of the association between the retriever and the target stimulus. However, the association between the retriever and the US was extinguished much faster and deeper with an intermediate ϑ, increasing the discrepancy in the level of activation of the mediated extinction cue and the US as the training progressed. As a result, the asymptote was also reduced. With more extinction trials

taking place and contributing to the total outcome prediction and a small asymptote, extinction developed more readily.

With a sufficiently robust neutral stimulus association capable of generating a strong retrieval of the associated cue, moderately elevated values of the variable ϑ are expected to increase the magnitude of the mediated conditioning and mediated extinction effects. These phenomena are largely driven by simple error correction mechanisms. Conversely, latent inhibition simulations were less susceptible to changes in the spread of associative activation engendered by ϑ, mainly relying on attentional-related factors. Although no effect of ϑ was observed for mediated latent inhibition, it did alter normal latent inhibition. In addition, the mediated enhancement of latent inhibition evidenced that learned attentional mechanisms are also subject to associative mediation. If within-stimulus associations are taken into account, their associative memory representations can become a contributing factor to latent inhibition and be susceptible to associative activation reverberation.

The mediated conditioning simulations presented in this paper closely match the results reported by Fleming et al. (2022), supporting the initial proposal that the amplification in the reverberation of the activation throughout the network produced by an increase of ϑ may be a suitable computational simile of ketamine-induced effects on mediated learning. However, even though the treatment between the empirical study and the reported simulations differed, the source of the effect in the simulation can only be traced to the mediated conditioning phase. ϑ's effect targets associative retrieval; hence, de facto operates as a specific manipulation that affects mediated learning. ϑ alters the reverberation of indirect activations in the network. Whereas it has a substantial effect in scenarios where mediated learning is the main factor generating the outcome, it has no significant effect in standard conditioning settings –where associations are mainly formed between directly input stimuli. In these scenarios, ϑ merely increases the number of stimulus elements available for

association. Moreover, since the model computes stimulus predictive values as averages of competing elements' weights and the activity of directly input stimuli is higher than that of retrieved elements, the effect in standard conditioning is immaterial. Yet, the manipulation could have produced an effect on the strength of the neutral stimuli associations developed in the compound preexposure phase, and this difference could have potentially increased the magnitude of mediated conditioning. However, no evidence of it was observed. The strength of the neutral stimuli associations at the start of the mediated conditioning did not substantially differ. The extinction rate of the neutral associations during mediated conditioning did diverge, but the source and location of this effect precisely lie in the boost in the reverberation of the stimulus activation during mediated conditioning. Therefore, the simulations and DDA account of the results offer an alternative analysis of the observed mediated learning enhancement, in which associative memories (internal activation of the stimulus representation by other stimuli present or associatively retrieved) re-echo or resonate in the network and become susceptible to learning –allowing the formation of spurious associations. This analysis is not incompatible with Fleming et al. (2022) reasoning and can parallel it.

If these simulations can be taken as a computational analogy that contributes to explaining the mediated enhancements reported in the literature following the administration of drugs, like ketamine, capable of inducing psychosis-like symptoms in rodents (Fleming et al., 2022; Fry et al., 2020; Koh et al., 2018; Razoux et al., 2007), the DDA model could offer a detailed and precise associative mechanism of how memory representations interact and conflict with actual experience. Data from laboratory animals has long related the dopamine system with the pathophysiology of schizophrenia (e.g., Miller, 1976; Robbins, 1990). Additionally, it has been suggested that dopamine error can become a signal that fosters associations, including those involving neutral stimulus relationships (Millard et al.,

2022). Furthermore, research exists that posits that delusions and hallucinations may result from the abnormal or aberrant association formation (Jensen et al., 2008), and support has been produced for associative models of delusion formation (Corlett et al., 2007b) along the lines analyzed in this paper.

More naturalistic modifications of these types of fully connected associative models that operate under simple error correction processes could be able to describe the development of the symptoms at different stages, effectively becoming a computational instantiation of animal models of schizophrenia (Becker et al., 2003). They could also serve as a tool to analyze and perhaps discriminate between theoretical accounts that emphasize different attentional and learning processes. For example, the fact that the model predicts no substantial differences due to the manipulation of ϑ for mediated latent inhibition could help elucidate the source of differential effects of ketamine on attentional and error correction mechanisms. Latent inhibition is assumed to critically depend on attentional processes. However, associative models conceptualize attention as a learning rate, an associability parameter that varies as an indirect effect of the error correction mechanism, under the assumption of an iterative process of updating the prediction of an absent motivationally relevant cue. Thus, altering the spread of associative activation of an assumed weak stimulus representation under these conditions may be insufficient to produce any noticeable effect. In these scenarios, perceptual expectations may play a more determinant role. For instance, drug-perceptual properties may act as an additional strong outcome that interferes with the putative predictor-outcome association.

Despite the significant advantages that DDA and other computational models provide, they rely on hand-crafted features. In the future, such models should be complemented with further developments in learning theory that incorporate a more naturalistic representation of the world (Mondragón et al., 2017) that, in turn, will lead to a

better understanding of the processes governing the intricate relationship between associative memories, attention, and learning. Until then, computational analyses, such as the one introduced in this paper, could offer some insight into clinical research and treatment. However, I would like to finish with a cautionary note: we must not mistake computational models for psychological theories. Computational models are necessary simplifications of reality. As formal instantiations of pre-existing psychological theories, their implementation can serve us to, firstly, precisely define and refine such theories, making them accountable, and, secondly, their algorithmic descriptions allow us to execute calculations rapidly and, more crucially, accurately but on their own, they do not embed the corpus of knowledge and intuitions of the community, nor are they necessarily guided by empirical evidence. (Alonso & Mondragón, 2012).

REFERENCES

Aitken, M. R., Larkin, M. J., & Dickinson, A. (2001). Re-examination of the role of within-compound associations in the retrospective revaluation of causal judgements. *Quarterly Journal of Experimental Psychology*, *54*, 27–51. http://dx.doi.org/10.1080/02724990042000029

Alonso, E., & Mondragón, E. (2012). Uses, Abuses and Misuses of Computational Models in Classical Conditioning. In N. Rußwinkel, U. Drewitz, J. Dzaack, H. van Rijn and Frank Ritterthe (Eds.), *Proceedings 11th International Conference on Cognitive Modeling (ICCM-12)* (pp. 96-100). Universitaetsverlag der TU Berlin. https://doi.org/10.14279/depositonce-3176

Baruch, I., Hemsley, D. R., & Gray, J. A. (1988). Differential performance of acute and chronic schizophrenics in a latent inhibition task. *The Journal of Nervous and Mental Disease*, *176*(10), 598–606.

Becker, A., Peters, B., Schroeder, H., Mann, T., Huether, G., & Grecksch, G. (2003). Ketamine-induced changes in rat behaviour: a possible animal model of schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *27*(4), 687-700. https://doi.org/10.1016/S0278-5846(03)00080-0

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114*(1), 80-99. https://doi.org/10.1037/0033-2909.114.1.80

Brandon, S. E., Vogel, E. H., & Wagner, A. R. (2000). A componential view of configural cues in generalization and discrimination in Pavlovian conditioning. *Behavioural Brain Research*, *110*(1-2), 67-72. https://doi.org/10.1016/S0166-4328(99)00185-0

Broadbent, D. (1958). *Perception and Communication.* Pergamon Press. https://doi.org/10.1037/10037-000

Corlett P.R., Honey G.D., & Fletcher, P.C. (2007a). From prediction error to psychosis: ketamine as a pharmacological model of delusions. *Journal of Psychopharmacology*, *21*(3), 238-252. https://doi.org/10.1177/0269881107077716

Corlett, P.R., Murray, G.K., Honey, G.D., Aitken, M.R., Shanks, D.R., Robbins, T.W., Bullmore, E.T., Dickinson, A., & Fletcher, P.C. (2007b). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain*, *130* (9), 2387-400. https://doi.org/10.1093/brain/awm173

Corlett, P.R, & Schoenbaum, G. (2021). Leveraging Basic Science for the Clinic-From Bench to Bedside. *JAMA Psychiatry*, *1*; 78(3):331-334. https://doi.org/10.1001/jamapsychiatry.2020.3656

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology B*, *37*, 397–416.

Dwyer, D.M. (2001). Mediated conditioning and retrospective revaluation with LiCl then flavour pairings. *The Quarterly Journal of Experimental Psychology B*, *54*(2),145-65. https://doi.org/10.1080/713932750

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological sciences*, *278*(1718), 2553–2561. https://doi.org/10.1098/rspb.2011.0836

Fleming, L.M., Jaynes, F.J.B., Thompson, S.L., Corlett, P. R., & Taylor, J. R. (2022). Targeted effects of ketamine on perceptual expectation during mediated learning in rats. *Psychopharmacology*, *239*, 2395–2405. https://doi.org/10.1007/s00213-022-06128-2

Fry, B.R., Russell, N., Gifford, R., Robles, C.F., Manning, C.E., Sawa, A., Niwa, M., & Johnson, A.W. (2020). Assessing Reality Testing in Mice Through Dopamine-Dependent Associatively Evoked Processing of Absent Gustatory Stimuli. *Schizophrenia Bulletin*, *46*(1), 54-67. https://doi.org/10.1093/schbul/sbz043

Granger, K.T., Talwar, A., & Barnett, J. H. (2020). Latent inhibition and its potential as a biomarker for schizophrenia. *Biomarkers in Neuropsychiatry*, 3, 1-6. https://doi.org/10.1016/j.bionps.2020.100025

Gray, J., Feldon, J., Rawlins, J., Hemsley, D., & Smith, A. (1991). The neuropsychology of schizophrenia. *Behavioral and Brain Sciences*, *14*(1), 1-20. https://doi.org/10.1017/S0140525X00065055

Hall, G. (1996). Learning about associatively activated stimulus representations: Implications for acquired equivalence and perceptual learning. *Animal Learning & Behavior*, *24*, 233-255. https://doi.org/10.3758/BF03198973

Haselgrove, M., & Hogarth, L. (Eds.). (2011). *Clinical Applications of Learning Theory*. Psychology Press. https://doi.org/10.4324/9780203803509

Hebb, D.O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons. https://doi.org/10.4324/9781410612403

Holland, P. C. (1981). Acquisition of representation-mediated conditioned food aversion. *Learning and Motivation*, *12*(1), 1-18. https://doi.org/10.1016/0023-9690(81)90022-9

Holland, P. C. (1983). Representation-mediated overshadowing and potentiation of conditioned aversions. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(1), 1–13. https://doi.org/10.1037/0097-7403.9.1.1

Holland, P.C. (1990). Event representation in Pavlovian conditioning: Image and action. *Cognition*, *37*(1-2), 105-131. https://doi.org/10.1016/0010-0277(90)90020-K

Honey, R. C., & Hall, G. (1989). Attenuation of latent inhibition after compound preexposure: Associative and perceptual explanations. *The Quarterly Journal of Experimental Psychology Section B*, *41*(4b), 355-368. https://doi.org/10.1080/14640748908401201

Jensen, J., Willeit, M., Zipursky, R., Savina, I., Smith, A.J., Menon, M., Crawley, A.P., & Kapur, S. (2008). The Formation of Abnormal Associations in Schizophrenia: Neural and Behavioral Evidence. *Neuropsychopharmacology*, *33*, 473–479. https://doi.org/10.1038/sj.npp.1301437

Kaplan, O., & Lubow, R. E. (2011). Ignoring irrelevant stimuli in latent inhibition and Stroop paradigms: the effects of schizotypy and gender. *Psychiatry Research*, *186*(1), 40–45. https://doi.org/10.1016/j.psychres.2010.07.025

Kapur, S. (2004). How antipsychotics become anti-'psychotic'—from dopamine to salience to psychosis. *Trends in Pharmacological Sciences*, *25*(8), 402 – 406. https://doi.org/10.1016/j.tips.2004.06.005

Koh, M.T., Ahrens, P.S., & Gallagher, M. (2018) A greater tendency for representation mediated learning in a ketamine mouse model of schizophrenia. *Behavioral Neuroscience, 132*(2), 106–113. https://doi.org/10.1037/bne0000238

Koh, M.T., & Gallagher, M. (2020). Using internal memory representations in associative learning to study hallucination-like phenomenon, *Neurobiology of Learning and Memory*, *175*, 107319. https://doi.org/10.1016/j.nlm.2020.107319

Kokkola, N., Mondragón, E., and Alonso, E. (2018). *Double Error Dynamic Asymptote (DDA) Model Simulator* (Version 1) [Computer software]. St. Albans, UK: CAL-R. Available at https://www.cal-r.org/index.php?id=DDA-sim and executables at https://github.com/cal-r/DDA_model

Kokkola, N., Mondragón, E., & Alonso, E. (2019). A double error dynamic asymptote model of associative learning. *Psychological Review*, *126*(4), 506-549. https://doi.org/10.1037/rev0000147

Kutlu, M.G., & Schmajuk, N.A. (2012). Solving Pavlov's puzzle: Attentional, associative, and flexible configural mechanisms in classical conditioning. *Learning & Behavior, 40*, 269–291. https://doi.org/10.3758/s13420-012-0083-5

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *The Quarterly Journal of Experimental Psychology B, 54*(4b), 311–352. http://dx.doi.org/10.1080/02724990143000072

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin, 142*(10), 1111–1140. https://doi.org/10.1037/bul0000064

Leung, H. T., Killcross, A. S., & Westbrook, R. F. (2011). Additional exposures to a compound of two preexposed stimuli deepen latent inhibition. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(4), 394. https://doi.org/10.1037/a0023714

Lewis, D. E., O'Reilly, M. J., Khuu, S. K., & Pearson, J. (2013). Conditioning the Mind's Eye: Associative Learning With Voluntary Mental Imagery. *Clinical Psychological Science*, *1*(4), 390–400. https://doi.org/10.1177/2167702613484716

Lubow, R. E., & Weiner, I. (Eds.). (2010). Latent inhibition: Cognition, neuroscience and applications to schizophrenia. Cambridge University Press. https://doi.org/10.1017/CBO9780511730184

Marcus, D.J., & Bruchas, M.R. (2021). Where ketamine and dopamine collide. *Elife*, *17*;10:e70148. https://doi.org/10.7554/eLife.70148

McLaren, I.P.L. (2011). APECS: An adaptively parameterized model of associative learning and memory. In E. Alonso, & E. Mondragón (Eds.), *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications* (pp. 145-164). IGI Global. https://doi.org/10.4018/978-1-60960-021-1.ch007

McLaren, I.P.L., Kaye, H., & Mackintosh, N.J. (1989). An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In R.G.M. Morris (Ed.), *Parallel Distributed Processing - Implications for Psychology and Neurobiology* (pp. 102-130). Oxford University Press.

McLaren, I.P.L., & Mackintosh, N.J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning and Behavior*, *38*(3), 211-246. https://doi.org/10.3758/BF03200258

Millard, S.J., Bearden, C.E., Karlsgodt, K.H., & Sharpe, M.J. (2022). The prediction-error hypothesis of schizophrenia: new data point to circuit-specific changes in dopamine activity. *Neuropsychopharmacology*, *47*, 628–640. https://doi.org/10.1038/s41386-021-01188-y

Miller, R. (1976). Schizophrenic psychology, associative learning and the role of forebrain dopamine. *Medical Hypotheses*, *2*(5), 203–211. https://doi.org/10.1016/0306-9877(76)90040-2

Miller, R. R., Kasprow, W. J., & Schachtman, T. R. (1986). Retrieval Variability: Sources and Consequences. *The American Journal of Psychology*, *99*(2), 145–218. https://doi.org/10.2307/1422275

Miller, R.R., & Matzel, L.D. (1988). The Comparator Hypothesis: A Response Rule for The Expression of Associations. *Psychology of Learning and Motivation*, *22*, 51-92. https://doi.org/10.1016/S0079-7421(08)60038-9

Miller, R. R., & Witnauer, J. E. (2016). Retrospective revaluation: The phenomenon and its theoretical implications. *Behavioural Processes*, *123*, 15–25. http://dx.doi.org/10.1016/j.beproc.2015.09.001

Mitchell, C., De Houwer, J., & Lovibond, P. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183-198. http://dx.doi.org/10.1017/S0140525X09000855

Mondragón, E., Alonso, E., & Kokkola, N. (2017). Associative learning should go deep. *Trends in Cognitive Sciences, 21*(11), 822-825. http://dx.doi.org/10.1016/j.tics.2017.06.001

Mondragón, E., Gray, J., & Alonso, E. (2013). A Complete Serial Compound Temporal Difference Simulator for Compound stimuli, Configural cues and Context representation. *Neuroinformatics*, *11*(2), 259–261. https://dx.doi.org/10.1016/j.cmpb.2013.01.016

Mondragón, E., Gray, J., Alonso, E., Bonardi, C., & Jennings, D. J. (2014). SSCC TD: A Serial and Simultaneous Configural-Cue Compound Stimuli Representation for Temporal Difference Learning. *PLOS ONE*, *9*(7), e102469. https://dx.doi.org/10.1371/journal.pone.0102469

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Rascle, C., Mazas, O., Vaiva, G., Tournant, M. Raybois, O., Goudemand, M., & Thomas, P. (2001). Clinical features of latent inhibition in schizophrenia. *Schizophrenia Research*, *51*(2–3), 149-161. https://doi.org/10.1016/S0920-9964(00)00162-6

Razoux, F., Garcia, R., & Léna, I., (2007). Ketamine, at a Dose that Disrupts Motor Behavior and Latent Inhibition, Enhances Prefrontal Cortex Synaptic Efficacy and Glutamate Release in the Nucleus Accumbens. *Neuropsychopharmacology*, *32*, 719–727. https://doi.org/10.1038/sj.npp.1301057

Rescorla, R. A. (1982). Simultaneous second-order conditioning produces S-S learning in conditioned suppression. *Journal of Experimental Psychology: Animal Behavior Processes, 8*(1), 23–32. https://doi.org/10.1037/0097-7403.8.1.23

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F.

Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). Appleton Century Crofts.

Robbins, T.W. (1990). The Case of Frontostriatal Dysfunction in Schizophrenia, *Schizophrenia Bulletin*, *16*(3), 391-402. https://doi.org/10.1093/schbul/16.3.391

Schachtman, T. R., & Reilly, S.S. (Eds.). (2011). *Associative Learning and Conditioning Theory: Human and Non-Human Applications*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199735969.001.0001

Schmajuk, N., & Alonso, E. (Eds.). (2012). Special issue on Computational Models of Associative Learning. *Learning & Behavior*, *40*.

Schmajuk, N. A., Lam, Y.-W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes, 22*(3), 321–349. https://doi.org/10.1037/0097-7403.22.3.321

Shanks, D., (2007). Associationism and cognition: Human contingency learning at 25. *The Quarterly Journal of Experimental Psychology*, *60*(3), 291–309. https://doi.org/10.1080/17470210601000

Shepard, R.N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking. and dreaming. *Psychological Review*, *91*(4), 417-447. https://doi.org/10.1037/0033-295X.91.4.417

Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In J. D. Moore, & J. F. Lehman (Eds.), *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 355–378). Erlbaum.

Thorn, P. D. (2021). The stimulus-to-perception connection: a simulation study in the epistemology of perception. *Synthese*, 199, 551–578. DOI:10.1007/s11229-020-02677-1

Tolman, E.C., & Brunswik, E. (1935). The organism and the causal texture of the environment. *Psychological Review*, *42*, 43-77. https://doi.org/10.1037/h0062156

Treisman, A., (1964). Selective attention in man. *British Medical Bulletin*, *20*, 12-16.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear, & R. R. Miller (Eds.), *Information Processing in Animals: Memory Mechanisms* (pp. 5–47). Erlbaum. https://doi.org/10.4324/9781315798820

Wagner A. R., & Brandon S. E. (2001). A Componential Theory of Pavlovian Conditioning. In R.R. Mowrer, & S.B. Klein (Eds.), *Handbook of Contemporary Learning Theories* (pp. 23–64). Erlbaum. https://doi.org/10.4324/9781410600691

Ward-Robinson, J., & Hall, G. (1996). Backward sensory preconditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 22*(4), 395–404. https://doi.org/10.1037/0097-7403.22.4.395

Table 1: Model fixed parameters. For a description of these parameters and their application within the model formalism, I refer the reader to the original paper (Kokkola et al. 2019).

Element/stimulus representation	CV	20
	Curve right skew (k)	2
	Set size	10
	Shared elements proportion	0.05
Context activation discount	CV element activation $\left(g^2 \right)$	2
Memory discounts	Backward discount (b)	1
	Eligibility Trace discount (γ)	0.998
	Eligibility exponent (z)	2/3

Table 2: Mediated Conditioning (MC), Mediated Extinction (ME) and Mediated Latent Inhibition (MLI) designs. Different values of the associative activation discount ϑ: 0.95, 0.975, and 1.05 were used for Group Standard, Group Raised, and Group High, respectively, for each phenomenon.

Mediated Conditioning (MC)	Phase 1: Compound Preexposure	Phase 2: MC	Phase 3: MC Test	Phase 4: DC Test
	AB; XY	B+; Y	A; X	B; Y

Mediated Extinction (ME)	Phase 1: Compound Preexposure	Phase 2: MC	Phase 3: ME	Phase 4: ME Test
	AB; XY	A+; X+	B	A; X

Mediated Latent Inhibition (MLI)	Phase 1: Compound Preexposure	Phase 2: MLI Preexposure	Phase 3: Conditioning (MLI Test)
	AB; XY	B	A+; X+; C+

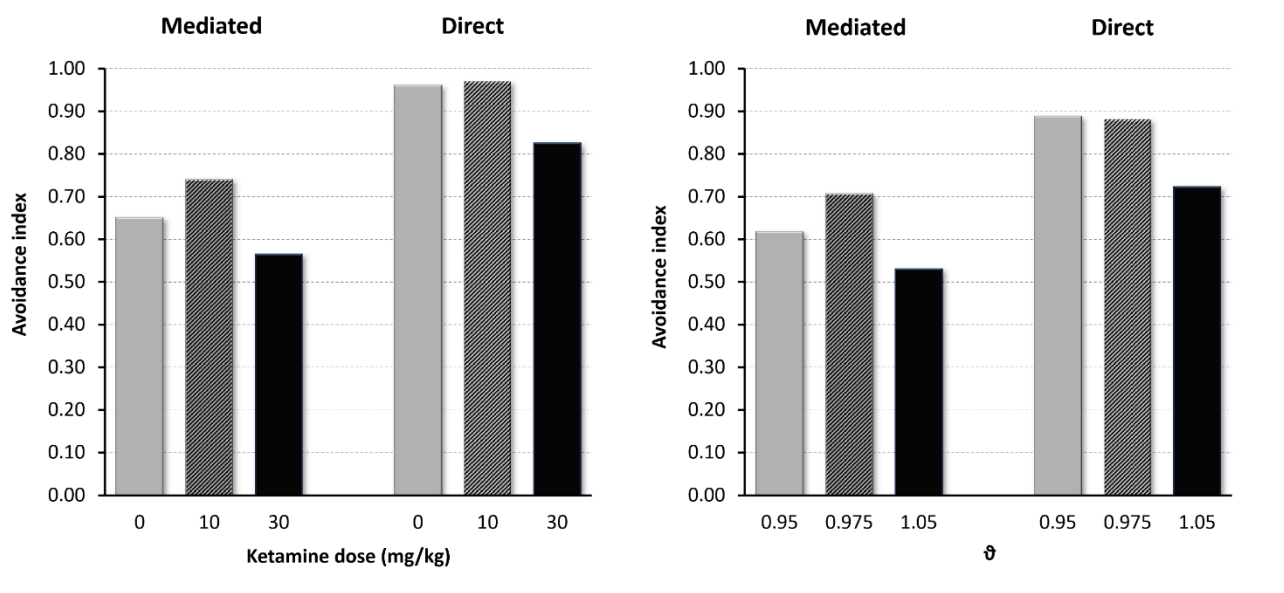

Figure 1: Simulation Set 1: Mediated Conditioning. The left panel is an adaptation of Fleming et al.'s. (2020) results (original measure) and displays a ratio of consumption avoidance per dose of ketamine, 0, 10, and 30 mg/kg. The right panel shows corresponding simulated results per ϑ running value. The index avoidance for the Mediated (MC Test, left columns) and Direct (DC Test, right columns) conditioned stimulus is presented on each panel.

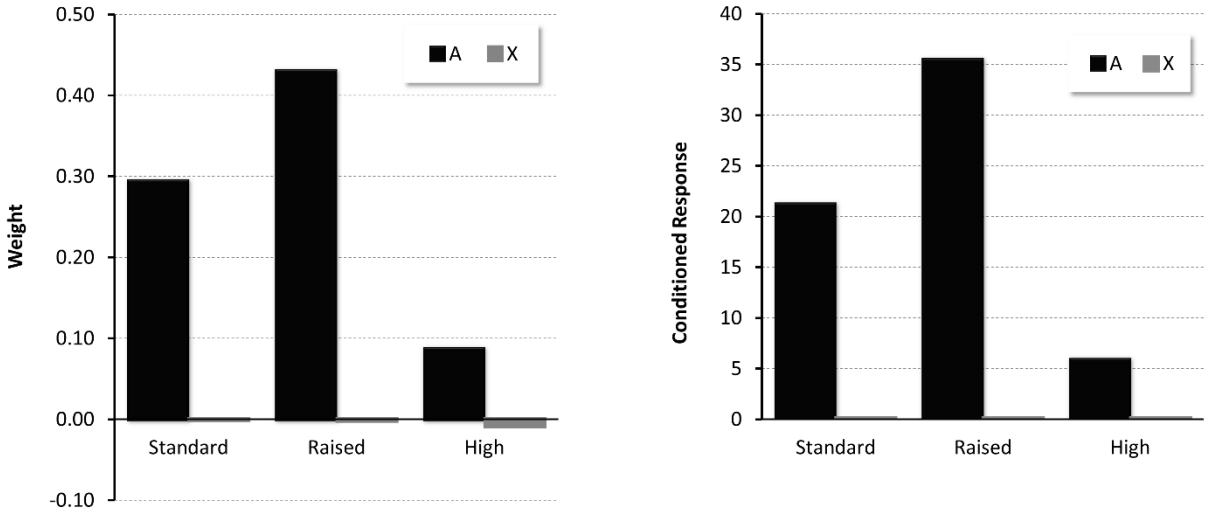

Figure 2: Simulation Set 1: Mediated Conditioning. Simulated MC Test results per cue and group. The left panel displays the weights of the stimulus-US associations for the mediated cue A (black) and directly activated cue X (grey) in Group Standard, Group Raised, and Group High. The right panel shows the corresponding predicted conditioned response.

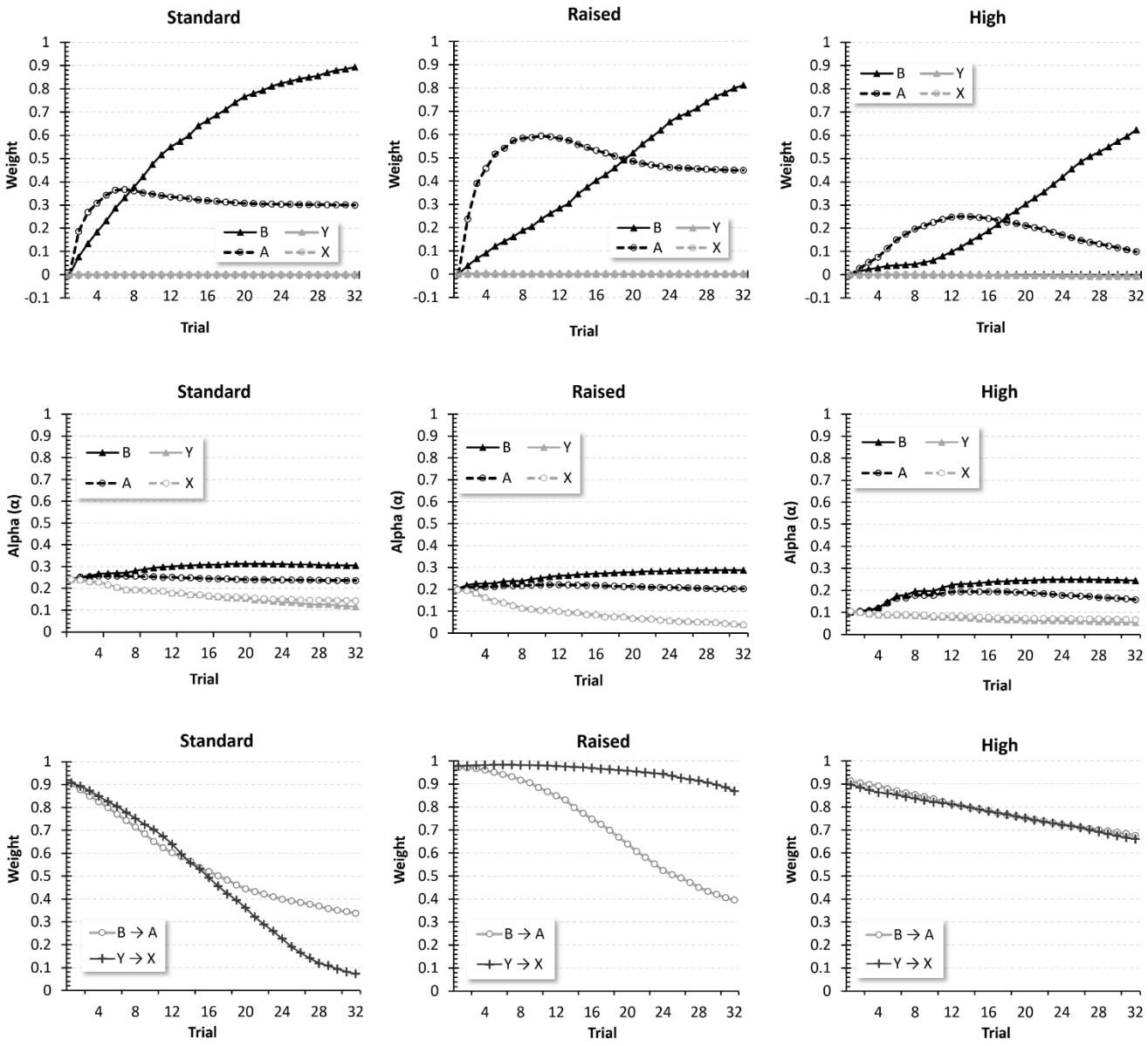

Figure 3: Simulation Set 1: Mediated Conditioning. The top panel shows the cue-US links' weights during Phase 2 (MC) for cues directly input (solid line), stimuli B (black, filled triangle) and Y (grey, filled triangle), and cues retrieved (dotted line), stimuli A (black, open circle) and X (grey, open circle), across trials. The middle panel shows the corresponding α values for stimuli B, Y, A, and X. The bottom panel displays the weights of the B-A association (grey, white circle) and Y-X association (black, cross) per trial. Each panel level presents the results for Group Standard (left panel), Group Raised (center panel), and Group

High (right panel).

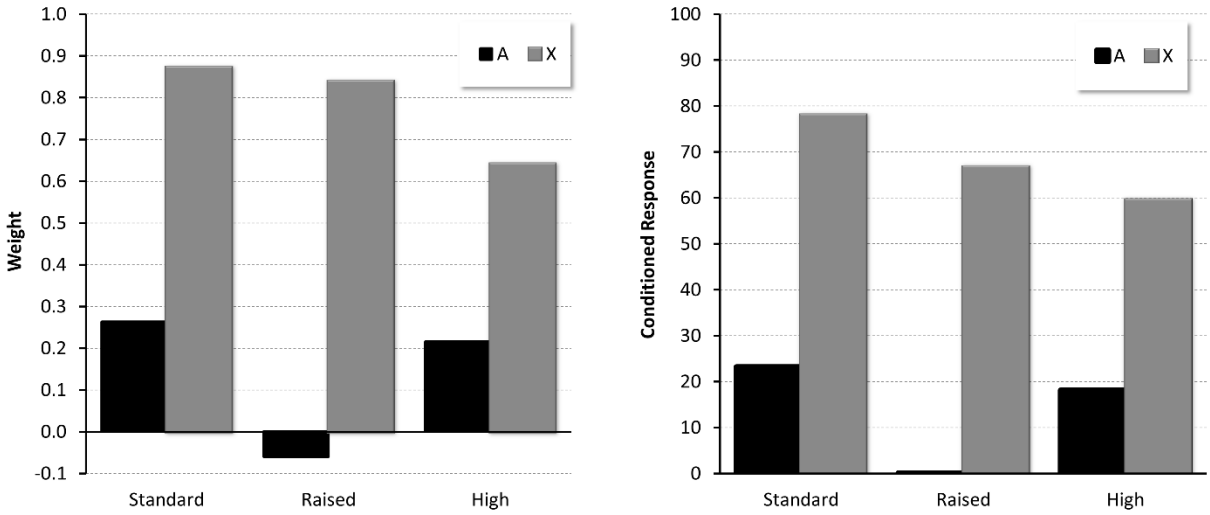

Figure 4: Simulation Set 2: Mediated Extinction. Phase 4 ME Test results per cue and group. The left panel displays the weights of the stimulus-US associations for the mediated cue A (black) and directly activated cue X (grey) in Group Standard, Group Raised, and Group High. The right panel shows the corresponding predicted conditioned response.

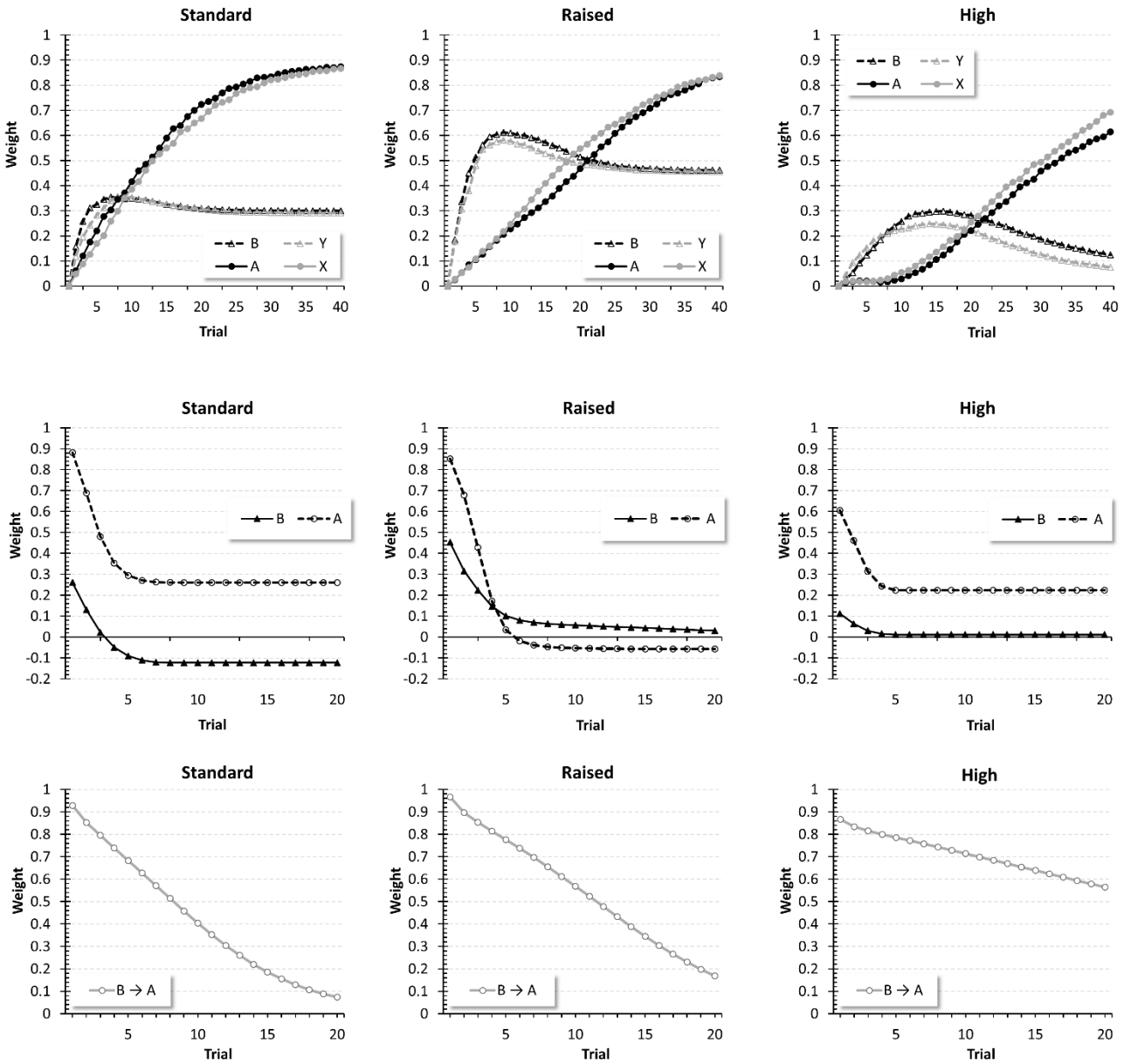

Figure 5: Simulation Set 2: Mediated Extinction. Simulated results for phases 2 (MC) and 3 (ME). The top panel displays cue-US links' weights during Phase 2 (MC) for the retrieved cues (dotted line), stimuli B (black, filled triangle) and Y (grey, filled triangle), and the direct input cues (solid line), stimuli A (black, open circle) and X (grey, open circle), across trials. The middle panel shows the Phase 3 (ME) weights for the directly input (solid line) stimuli B (black, filled triangle) and the retrieved (dotted line) stimuli A (black, open circle) links to the US across trials. The bottom panel displays the Phase 3 (ME) weights for the B-

A association (grey, white filled circle) per trial. Each panel level presents the results for Group Standard (left panel), Group Raised (center panel), and Group High (right panel).

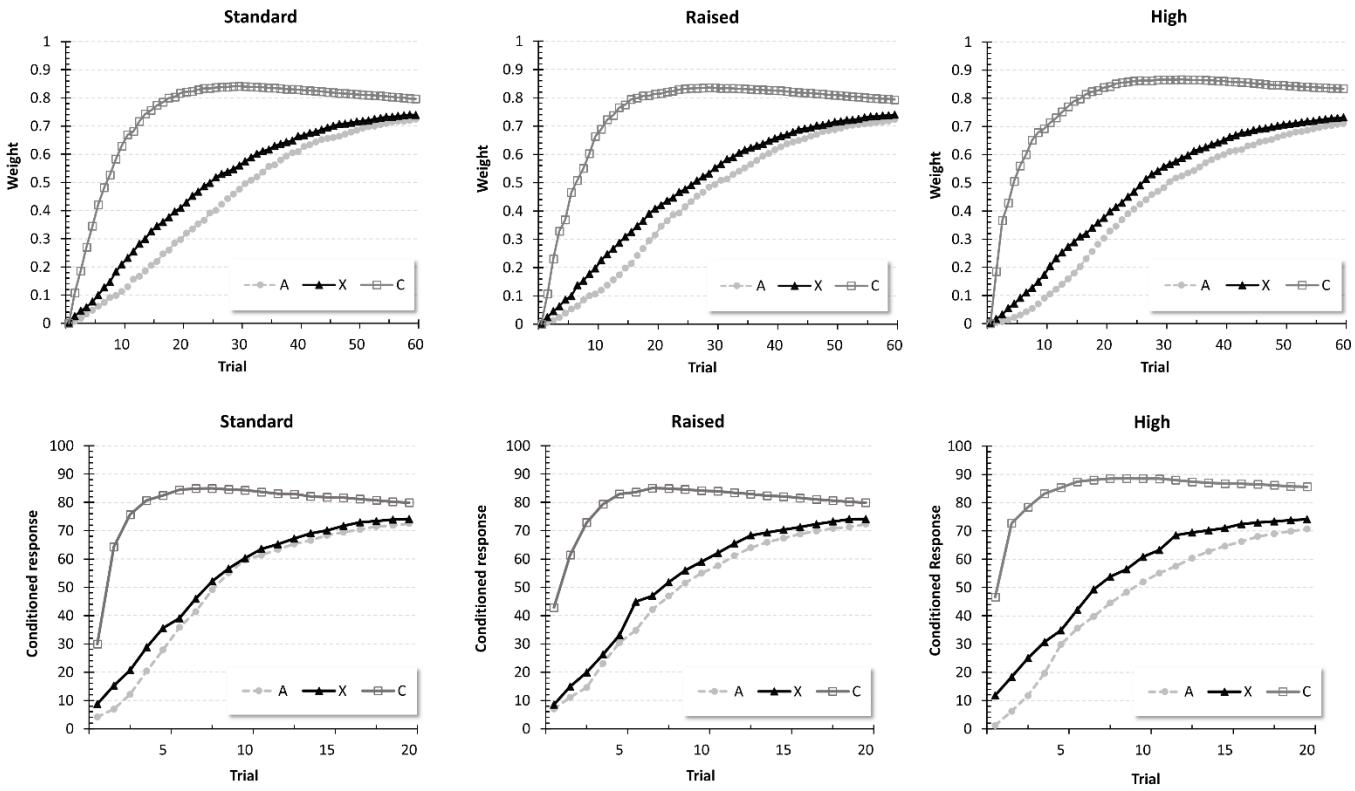

Figure 6: Simulation Set 3: Mediated Latent Inhibition. The top panel displays the weights of the association between stimulus A (clear-grey dotted line, filled circle), X (black solid line, filled triangle), and C (dark-grey solid line, open square) during Phase 3 MLI Test in Group Standard (left panel), Group Raised (center panel) and Group High (right panel). The bottom panel presents the corresponding predicted conditioned response.

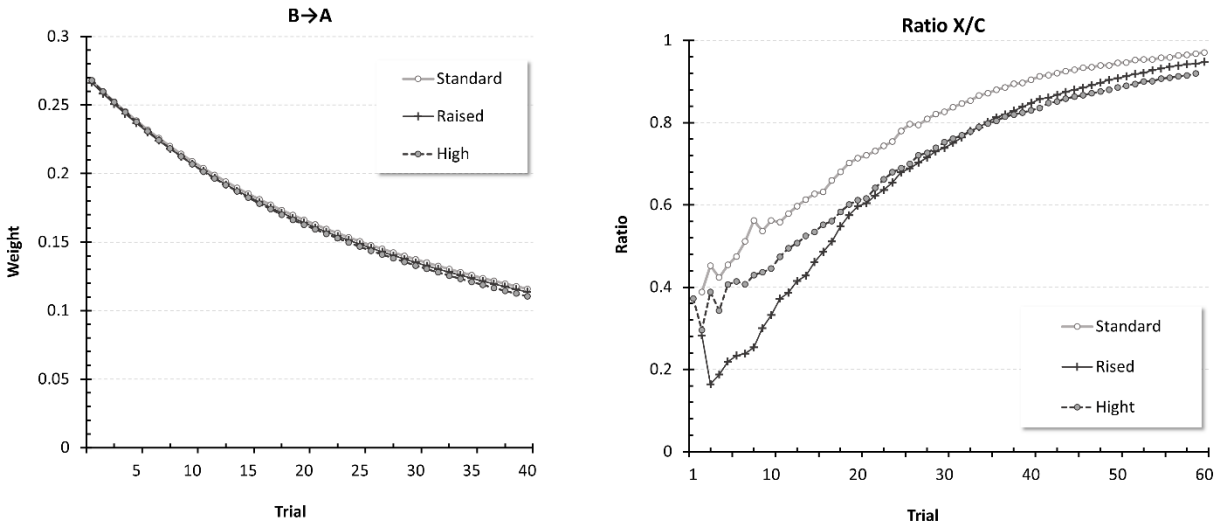

Figure 7: Simulation Set 3: Mediated Latent Inhibition. The left panel shows the weights of the B-A association during Phase 2 (MLI Preexposure); the right panel plots a ratio of weights between the X-US and the C-US links during Phase 3: Conditioning (MLI test) per Group Standard (solid grey line, white filled circle), Group Raised (solid black line, cross) and Group High (dotted black line, grey filled circle) per trial.

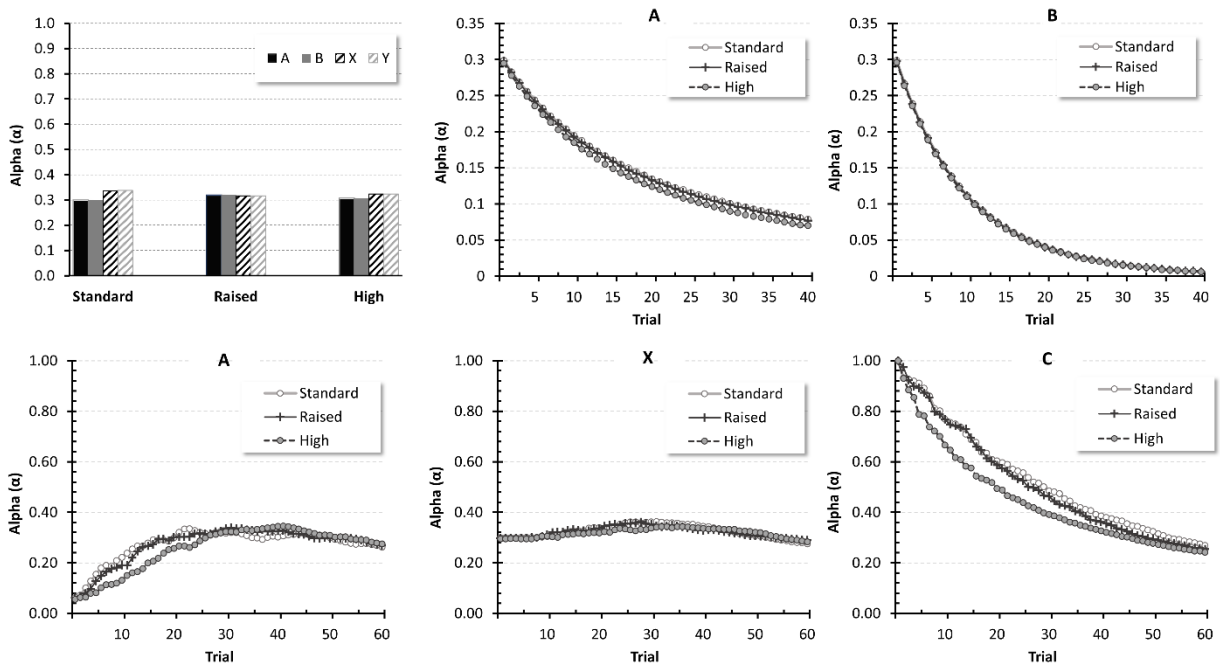

Figure 8: Simulation Set 3: Mediated Latent Inhibition. Simulated associability (α) values. The top-left panel displays the values per stimuli, A, B, X, and Y, in consecutive bars, during the last trial of Phase 1 (Compound Preexposure) per Group Standard (left), Raised (center), and High (right). The top-center and right panels show the decay in α of A and B, respectively, throughout Phase 2 (MLI Preexposure) per Group Standard (solid grey line, white filled circle), Raised (solid black line, cross) and High (dotted black line, grey filled circle). The bottom panel displays α values during Phase 3 (MLI Test) for stimuli A (left panel), X (center panel), and C (right panel) per group.