This chapter was originally published in the book *Computational Psychiatry*. The copy attached is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research, and educational use. This includes without limitation use in instruction at your institution, distribution to specific colleagues, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution’s website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier’s permissions site at:

http://www.elsevier.com/locate/permissionusematerial


ISBN: 9780128098257

Copyright © 2018 Elsevier Inc. All rights reserved.

Academic Press
CHAPTER 9

Modeling Negative Symptoms in Schizophrenia

Matthew A. Albrecht¹, ², James A. Waltz¹, Michael J. Frank³, James M. Gold¹

¹ University of Maryland School of Medicine, Baltimore, MD, United States; ² Curtin University, Perth, WA, Australia; ³ Brown University, Providence, RI, United States

OUTLINE

9.1 Introduction: Negative Symptoms in Schizophrenia 220
9.2 Dopamine Systems and Prediction Errors 222
9.3 Modeling in Reward-Related Decision Tasks 224
9.4 Probabilistic Stimulus Selection—Combined Actor-Critic/Q-Learning 224
  9.4.1 Rationale 224
  9.4.2 Q-Learning 227
  9.4.3 Actor-Critic Model 228
  9.4.4 Combined Actor-Critic/Q-Learning 228
  9.4.5 Findings 229
9.5 Time Conflict—Temporal Utility Integration Task 230
  9.5.1 Rationale 230
  9.5.2 Time Conflict Model 232
  9.5.3 Findings 233
9.6 Pavlovian Bias—Extended Q-Learning
9.6.1 Rationale 235
9.6.2 Pavlovian Bias Model 236
9.6.3 Findings 237

9.7 Direct Addition of Working Memory to Reinforcement Learning Models
9.7.1 Rationale 240
9.7.2 Reinforcement Learning and Working Memory Model 240
9.7.3 Findings 241

9.8 Summary 242
9.9 Conclusion 244
References 244

How modeling can inform reward-related decision tasks to uncover the computational bases of decision-making impairments in schizophrenia in relation to negative symptoms, which has implications for understanding motivational disturbances in other affective disorders such as major depression.

9.1 INTRODUCTION: NEGATIVE SYMPTOMS IN SCHIZOPHRENIA

Negative symptoms are a cluster of self-reports and behaviors characterized by the reduction of normal activity or emotion, the origin of which remains unknown. The expressed symptoms superficially overlap with much of the classic depressive disorder phenotype and include motivational deficits, anhedonia, avolition, and reduced affective expression. These symptoms in the context of schizophrenia have seen a steadily increasing research focus, including more direct attempts at investigating treatment options in a hope to alleviate the poor functional outcome (Norman et al., 2000; Evans et al., 2004), poor quality of life (Katschnig, 2000), and low rate of recovery (Strauss et al., 2010). Unfortunately, there is currently no Food and Drug Administration approved drug for the treatment of negative symptoms. This is partly due to a function of our poor understanding of the etiology of negative symptoms, which is centered on clinical observations and phenomenological interrogation.

Some avenues of phenomenological interrogation may have led us astray, having a significant impact on our ability to understand negative symptoms and impaired our ability to develop valid preclinical models,
highlighting the need for more formal models of symptom etiology. As an example of where we may have been led astray, it would appear reasonable at first to assume that the negative symptom anhedonia represents an emotional construct of diminished pleasure and therefore diminished in the moment pleasure would be reduced in people exhibiting this negative symptom. However, while patients may report less pleasure relative to controls when reporting on noncurrent feelings, they report similar levels of in the moment/current positive emotion similar to healthy subjects (Kring and Moran, 2008; Cohen and Minor, 2010). It is the former that appears to drive ratings on clinical symptom expression scales (Strauss and Gold, 2012), while we have developed much of our beliefs and animal models on the latter (e.g., sucrose preference test (Papp et al., 1991)). As a consequence, more recent thoughts on anhedonia suggest that it is likely due to a combination of downgrading the hedonic experience of past events and predictions about future events to be more pessimistic than may be the case (Strauss and Gold, 2012). While patient reports of the hedonic quality of past experiences and subsequent behaviors may contribute to nonadaptive behaviors in the future, they would still appear to be entirely rational taking into account a likely history of patient experience. That is, people with schizophrenia are likely to have significantly more negative experiences over their lifetime compared with others, especially if their experiences are interrupted by, or are predicted to be interrupted by, paranoid thoughts or malicious voices.

The refined understanding of negative symptoms was formed through a more extensive and detailed examination of clinical and patient reports, both important sources of information. However, it can sometimes be difficult to advance mechanisms using only subjective self-report sources, highlighting the need to combine more modern perspectives on the clinical/experiential aspects of negative symptoms with a more detailed/objective assessment of functionality based in formally quantified behavioral paradigms. Our research has used behavioral assessments of reinforcement learning (RL) to examine competing models of negative symptom formation. Broadly, patients with schizophrenia have repeatedly shown performance impairments in RL tasks. Intuitively, it can be seen how this could lead to deficits in goal directed action that underpin negative symptoms via straightforward impairments in learning about rewards and then acting on them. More formally, one explanation for these deficits and the propensity to develop negative symptoms could be because there is a failure in signaling reward prediction errors (PEs). A failure to adequately signal reward PEs could happen via two distinct mechanisms, either through a dampened PE response to reward or through noise in the reward PE system such that true reward PEs are drowned out by aberrant signaling (Maia and Frank, 2017). This idea is particularly appealing given the significant role for dopamine systems in
motivation and action. Furthermore, it is broadly accepted that dopamine contributes in some way to the clinical profile of positive symptoms of psychosis, and a common underlying neural substrate for schizophrenia possesses a seductive symmetry and parsimony. However, maladaptive value adjustment during RL can happen for any number of reasons besides faulty ventral-striatum reward prediction error signaling.

We will argue below that a critical cognitive pathology in schizophrenia is an inability to effectively and/or flexibly update representations of reward value over time. This specific deficit may more specifically lead to the types of failures in RL performance in schizophrenia, particularly in those patients that present with particularly severe negative symptoms. The key studies supporting this hypothesis were formulated using a computational psychiatry perspective in an attempt to isolate the core computational constructs/modules used by the brain to calculate, store, and act on value. These computational constructs are an abstraction of brain processes where it is assumed that an underlying neural network is associated with a particular calculation. Therefore, these computational models are designed to overlap with neural models thought to generate the expression of behavior. As elaborated below, this framework has led to alternative computational accounts that posit reward-learning deficits in schizophrenia that can arise from alterations within the core dopaminergic RL system itself (Maia and Frank, 2017), or from alterations in more cognitive prefrontal mechanisms that, in the healthy brain, are used to elaborate and enhance learning (Strauss et al., 2011; Gold et al., 2012; Collins et al., 2014; Collins et al., submitted). With a more refined characterization and quantification of the contributions of multiple systems to learning and decision-making, one can begin to design tasks to be more sensitive and specific to disentangle their contributions.

We start with a brief overview of dopamine’s role in representing reward value and its anatomical and functional interactions with frontal regions that form a basal-ganglia-cortico-striatal network, as an initial link from brain function to computational abstraction. Following from this, we will describe the application of computational models to RL tasks and their insights into schizophrenia.

### 9.2 DOPAMINE SYSTEMS AND PREDICTION ERRORS

Dopamine neurons in the mesolimbic system seem to play an important role in signaling the value of a stimulus. Increases in phasic firing of dopamine neurons occur in response to the unexpected delivery of a rewarding stimulus. The onset of firing can also be shifted from the rewarding stimulus to a stimulus predictive of that reward, as hypothesized by an extension of classical associative learning schemes to RL.
algorithms that span temporally extended tasks (Montague et al., 1996). Moreover, if the reward is omitted at the expected time of delivery after a presentation of a reward-predicting stimulus, dopamine neuron firing rate will be suppressed. This simple bidirectional response to value provided a foundation for the idea that dopamine neurons signal a valenced reward PE, their firing rate increasing in response to unexpected reward (or reward prediction), and reducing in response to reward omission, thereby signaling positive and negative PEs, respectively. The magnitude of phasic bursts and dips of DA are correlated with the magnitude of positive and negative PEs, respectively (Bayer et al., 2007; Hart et al., 2014). Dopamine signaled PEs go on to modulate decision-making/action selection by modulating synaptic plasticity of the circuits responsible for action/behavioral choices. Dopamine can also affect decision-making by modulating choice incentive directly, e.g., positive PEs that increase dopamine can result in a D₁-receptor system bias associated with Go-actions adjusting behavior toward riskier options but with higher potential value (Collins and Frank, 2014; Maia and Frank, 2017). Optimal action selection is supported by learning systems situated in frontal cortex, in particular the orbitofrontal cortex (OFC) and medial prefrontal cortex (mPFC). These frontal cortical regions appear to be responsible for calculating, maintaining, and weighing stimulus-action-value associations (Schoenbaum and Roesch, 2005; Frank and Claus, 2006; Fellows and Farah, 2007). Moreover, these frontal systems support rapid and flexible updating of stimulus-action-value calculations compared to the slower updating in the basal ganglia. This ability of OFC to weigh up multiple actions allows simultaneous evaluation of multiple action-outcome scenarios, aiding optimal decision-making. The key limitation here is that action-outcome scenarios are constrained within the capacity of working memory (WM). These two systems play interacting roles in RL and decision-making, complementing each other’s strengths and weaknesses. For example, the OFC is able to rapidly update stored values and can directly contrast several competing options, providing an additional top-down influence on decision-making. However, the system is resource-limited particularly by constraints on WM. By contrast, the striatal system contributes a more stable stimulus valuation, and actions are integrated in a more habitual manner. Consequently, the striatal system is not as resource-limited by WM capacity. However, this system can be less flexible to alterations in value and is limited by the actions it can select.

Of course, one of the most influential models of schizophrenia relates to the function of the dopamine system. Computational research has therefore followed this focus, attempting to formalize a relationship between aberrant dopamine signaled positive PEs and the expression of positive symptoms. Computational models of RL can potentially be informative about negative symptoms. The systems that we have discussed above are
integral components involved in motivation, pleasure seeking, and action selection; functions that clinically appear to underlie negative symptom expression. Computational modeling is therefore a potentially useful tool to tease apart the separable contributions of these multiple interacting systems (and components within these systems), which may offer key insights.

9.3 MODELING IN REWARD-RELATED DECISION TASKS

Computational modeling of RL tasks has at its core the concept of the reward PE; the difference between an outcome that was expected and an outcome that was experienced. Mathematically, this can be expressed as:

\[ \delta_t = r_t - V_t \]

where \( r_t \) is the experienced reward for trial \( t \), \( V_t \) is the expected reward value, and \( \delta_t \) is the reward PE. This core module will be a critical component of the models described below where we attempt to understand behavior on a trial-by-trial basis. Using the foundational concept of a reward PE, we can build models of increasing complexity to describe the behavioral responses to stimuli. These models can include any number of parameters depending on which computational principles are relevant (and can be identified) when the subject is confronted with a particular task. Of course, the parameters incorporated into the model will depend strongly on the structure of the task. Below we begin with a relatively simple stimulus-selection task that incorporates probabilistic rewards and punishments to shape action selection.

9.4 PROBABILISTIC STIMULUS SELECTION—COMBINED ACTOR-CRITIC/Q-LEARNING

9.4.1 Rationale

In a previous report, we had shown that people with schizophrenia demonstrated impaired acquisition in a simple probabilistic RL task (Waltz et al., 2007). The task required participants to select between two stimuli that were reinforced at either 80% (the optimal stimulus) or 20%. They also had to learn to select among two other pairings of stimuli with more probabilistic values (70/30 and 60/40). An example of the structure of this task is described in Fig. 9.1. This kind of task is learned relatively
quickly in control participants who learn to select the most rewarding stimulus of the various pairs. By contrast, patients with schizophrenia showed learning delays at the early stages of training. In a subsequent transfer phase, where the previously trained stimuli were presented to participants without feedback, both controls and patients were able to select the most rewarding stimulus compared with the least rewarding one. Patients did not, however, pick the high-value stimulus (that which would have elicited the largest number of positive PEs) over a neutral novel stimulus as reliably as controls, but they were able to match controls’ performance in avoiding the stimulus with lowest reward (that which would have elicited the largest number of negative PEs, and also likely the fewest number of experienced feedback events assuming they learned the task) stimulus with the same neutral novel stimulus.

It was suggested that this pattern of results indicated that patients did not have a problem with learning from negative PEs, they were only selectively impaired at learning from positive PEs. However, while
positive PE learning is one mechanism required for subsequent choice of
the more rewarding stimulus in this task, it is not the only one. Indeed,
while the original task was designed to test a computational model of
striatal dopamine function on positive and negative PE learning, subse-
quent computational analysis based on the same framework suggested
that even when positive learning is equivalent across groups, impaired
choice of the most rewarding stimulus can arise from a reduced repre-
sentation of the expected positive value of that action during action
selection (Collins and Frank, 2014). In turn, this deficiency can arise from
either degraded OFC representations of expected value (EV) (Frank and
Claus, 2006), or from reduced striatal DA signaling during the choice
period, leading to reduced choice incentive and a greater emphasis on
negative outcomes (Collins and Frank, 2014). Both mechanisms have been
posited to account for this selective reduction in choice of positively
valued options during learning in people with schizophrenia (Gold et al.,
2012; Huys et al., 2016). Moreover, patients received less exposure to the
high-value stimulus as they were slower to learn the association,
increasing the uncertainty of its value. Therefore, a reduced preference for
the high-value stimulus compared to controls may also be related to less
exposure making the positive EV more uncertain.

To assess some of these hypotheses, we implemented a similar probabi-
listic RL task that included reward and punishment contingencies to obtain
positive and negative reward PEs in both instrumental contexts. That is, in
addition to the above (probabilistic) reward versus no-reward condition, a
second condition was included where participants learned based on
(probabilistic) loss versus no-loss (loss-avoidance). Critically, the same zero
outcome would yield both positive and negative PEs depending on gain or
loss context. This modification addresses one of the critical confounds in the
above experiment; the conflation of reward with positive PE. The previous
experiment also did not model behavior to corroborate whether poor PE
signaling, or other factors contributing to RL performance, more explicitly
contributed to patients performance differences. Therefore, to further
examine whether positive PEs or deficits in maintaining EV were potential
reasons for patient-control differences, we used a combination of two
interacting RL models: a “Q-learning” model and an “Actor-Critic” model.
We begin by describing the Q-learning model, a simple two-parameter
model that keeps track of state-action value pairs, before introducing the
Actor-Critic model. Then the process of combining these two models into a
single model with two distinct learning modules will be described. Each of
these models is theoretically linked in some way with the basal ganglia
system and reward PEs, but each make different predictions about the na-
ture of representations used in reward-based decision-making.
9.4.2 Q-Learning

The Q in Q-learning represents the expected utility (Quality) of a state-action pairing. Q-learning models learn about state-action pairs, such that for any given state, multiple possible actions may be entertained. For example, for a given state (or stimulus) $X_1$, an action $Y_1$ possesses an EV 0.5, while action $Y_2$ may possess a value of 0.7. A series of states and possible actions yields a state-action matrix with the dimension of the rows and columns depending on the number of states and actions, respectively e.g.,

$$
\begin{array}{c|c|c}
\text{Action} & Y_1 & Y_2 & Y_j \\
\hline
X_1 & 0.5 & 0.7 & \ldots \\
X_2 & 0.2 & 0.4 & \ldots \\
X_i & \ldots & \ldots & \ldots \\
\end{array}
$$

The values within the state-action matrix naturally change over time depending on the past history of reinforcements and punishments, and are updated based on a weighted PE, such that:

$$
\delta_t = r_t - Q_t(s, a)
$$

$$
Q_{t+1}(s, a) = Q_t(s, a) + \alpha \delta_t
$$

where for trial $t$, $r_t$ gives the obtained reinforcement value for that particular trial (usually given as $+1$, $0$, or $-1$ for reward, neutral, and punishment, respectively), $\delta_t$ is the PE for the state-action pair at trial $t$, and $\alpha$ describes the learning rate or the speed in which the state-action pairings update.

For a given new stimulus, action selection for that particular state is made probabilistically based on contrasting the learned Q-values of each candidate action for that state. A transformation of the Q-values into an action probability is often obtained through a softmax function, transforming the Q-values into a probability between 0 and 1. These values reflect the probability of selecting that particular action for that state. The softmax function is given as:

$$
p(a|s) = \frac{\exp(\beta Q_t(s, a))}{\sum_{i=1}^{N_a} \exp(\beta Q_t(s, a_i))}
$$

where $\beta$ represents the temperature, determining the degree with which differences in Q-values are translated into a more deterministic choice.
9.4.3 Actor-Critic Model

The actor-critic model is a two-part model with many of the same features as the Q-learning model. One component, the “Critic” is an evaluator of the state’s value. The Critic learns the value of the stimulus without taking into account the possible actions. The second component, the “Actor,” is used for action selection and learns stimulus-response weightings for each state-action pair as a function of the critic’s evaluation. PEs are generated at the Critic level to update both the state value of the critic and the stimulus-response weights of the Actor. The model is formalized as follows, first the critic PE for each trial is obtained similar to the Q-learning model above for each stimulus (note not a stimulus-action pair):

$$\delta_t^{\text{Critic}} = r_t - V(s_t)$$

The critic PE $\delta_t^{\text{Critic}}$ then updates the stored value of the state representation modified by the critic’s learning rate $\alpha^{\text{Critic}}$:

$$V_{t+1}(s) = V_t(s) + \alpha^{\text{Critic}} \delta_t^{\text{Critic}}$$

The critic’s PEs is also used to update the actor’s response weighting matrix for that stimulus, by:

$$W_{t+1}(s,a) = W_t(s,a) + \alpha^{\text{Actor}} \delta_t^{\text{Critic}}$$

where $\alpha^{\text{Actor}}$ represents the Actor’s learning rate. Note also the difference between the stored representations of the critic’s value $V_t(s)$ and the Actor’s weights $W_t(s,a)$, where the former is only representing states and the latter is representing state-action pairings. As can be seen from the formulas, there are two separate learning rates, one for the critic and one for the actor, that represent the speed of updating of each system’s values and weights, respectively. The cognitive neuroscience literature has linked the ventral striatum to the critic and the dorsal striatum to the actor, both of which learn from predictions errors signaled by dopamine (O’Doherty et al., 2004). Finally, the action weights are normalized between 0 and 1 by:

$$W_{t+1}(a,s) = \frac{W_{t+1}(s,a)}{\sum_{i=1}^{N_t} W_{t+1}(s,a_i)}$$

9.4.4 Combined Actor-Critic/Q-Learning

The above two models represent two unique strategies for learning stimulus-action reinforcement and putatively represent two separate neurological systems. The Q-learning model selects actions by comparing Q-values of state-action pairings and then updates these using a delta rule.
on the state-action values directly. By contrast, in the actor-critic model, the critic evaluates the overall state without considering competing actions and then feeds this information to the actor that selects actions with strong stimulus-action weights (without directly evaluating the stimulus-action value). The Q-learning model is hypothetically linked to OFC that maintains value representations of the state-action pairs, whereas the actor-critic model is linked more strongly to the basal ganglia system. In the brain, these two systems interact to solve action selection each contributing their unique perspective to the problem at hand. We can model this interaction using a hybrid model, whereby each system is updated independently, with critic PEs and Q-learning PEs calculated for the separate systems and the final action selection is calculated by adding a mixing parameter that apportions the degree to which each system contributes to the final stimulus-action weight, by:

\[ q = (1 - mix) \times W_i(s,a) + mix \times Q_t(s,a) \]

where \( q \) is the final stimulus-action weight for all possible actions given a stimulus, \( mix \) describes the proportion of each model used, and \( W_i(s,a) \) and \( Q_t(s,a) \) are the actor-critic and Q-learning contributions, respectively, calculated as above. Other more elaborated actor-critic models separate the actor into two opponent processes, capturing asymmetric effects of dopamine on learning and choice based on positive versus negative incentives (Collins and Frank, 2014) and which can be applied to some of the schizophrenia literature (Maia and Frank, 2017), but for brevity we do not detail this model here.

### 9.4.5 Findings

Using this task we found that patients with high number of reported negative symptoms learned best when attempting to avoid punishments, while performing poorest when learning from rewards in a 90% contingency condition compared with both controls and low negative symptom (LNS) patients (Gold et al., 2012). Poor reward-learning performance was further shown in the transfer phase, where high negative symptom (HNS) patients were less likely to select the higher value stimulus, particularly when comparing a frequently winning stimulus with a frequently loss-avoiding stimulus. HNS patients essentially treated the accumulated positive PEs during the reward condition and the loss-avoidance condition with a more similar weight, whereas controls and LNS patients more definitively favored positive PEs from rewarding stimuli over loss-avoidance stimuli. Compared with controls and LNS patients, HNS patients also failed to select the stimulus with the highest EV in the easiest transfer condition comparing the frequently rewarded stimulus with the frequently punished stimulus. A failure to select the more obvious
optimal stimulus implicates a general failure in maintaining stimulus value, alternatively it could be due to the reduced exposure to the most frequently rewarded stimulus reducing the certainty surrounding its EV.

The modeling results yielded further information on the cause of the RL performance differences in HNS. In particular, differences between HNS patients and controls were seen in the mixing parameter that describes the ratio of Q-learning to Actor-Critic learning. The mixing proportion in controls was fit at a ratio of ~0.7 in favor of the Q-learning model, i.e., a greater proportion of their action selection used the model most related to the EV of the choice (thought to be represented in OFC) rather than just the history of PEs. By contrast, in patients with a high burden of negative symptoms, the mix between the Q-learning module and the Actor-Critic module was substantially lower at ~0.4. LNS patients sat between HNS and controls at ~0.6. A reduction in the contribution of the Q-learning component to RL in HNS patients implicates a possible mechanism for poor RL performance and its relationship to negative symptoms: a failure of top-down control by the OFC impairs maintenance of EV, particularly for stimuli associated with rewarding outcomes (not necessarily impaired for positive PEs).

### 9.5 TIME CONFLICT—TEMPORAL UTILITY INTEGRATION TASK

#### 9.5.1 Rationale

The probabilistic stimulus-selection task described above showed that patients seem to possess difficulties integrating a history of rewarding outcomes and selecting those stimuli over others. Difficulties specific to reward-learning may suggest deficits in striatal DA D1-receptor signaling. The D1 subsystem of the basal ganglia is linked with making actions in response to rewarding outcomes following phasic dopamine bursts. Therefore deficits in this system may be responsible for poor reward Go-learning in the HNS patients. Conversely, there appears to be intact responding to negative PE signaling in HNS patients, which is ostensibly related to the D2 receptor subsystem that is sensitive to phasic dopamine dips and actions within the NoGo pathway. The basal ganglia D1 and D2 subsystems also modulate response times. Following positive PEs and phasic dopamine bursts targeting the D1 subsystem, response times are reduced and Go-learning is enhanced. Following negative PEs and phasic dopamine dips targeting the D2 subsystem, response times slow and NoGo-learning is enhanced.

We can use this phenomenon to investigate the potential role of D1 and D2 systems during RL in people with schizophrenia using a slightly
different approach to the probabilistic selection task. We administered the temporal utility integration task (TUIT) that adjusts reward probabilities and magnitudes as a function of time elapsed (Moustafa et al., 2008). Fig. 9.2 illustrates the experimental procedure for the TUIT. Participants are presented with a clock face with a single arm, where the arm turns over the face of the clock in around 5 s. Participants are required to press a button at any time between the onset of movement of the clock arm and a complete rotation. Points are given to the participant depending on four conditions, where the EV (probability × magnitude) for each condition across the 5 s interval is defined as:

- Decreased expected value (DEV) = early responses yield a higher number of points than expected and should therefore lead to Go-learning/speeding. Faster responses gave more points on average.
Increased expected value (IEV) = early responses yield a small number of points, which should lead to NoGo-learning and slowed responses. Slower responses gave more points on average.

Constant expected value (CEV) = reward probability increases, while magnitude decreases as time elapses on the clock.

Constant expected value Reversed (CEVR) = reward probability decreases, while magnitude increases as time elapses on the clock.

Using these last two conditions which have constant EV over the clock face, we can determine relative biases in a person’s preference for reward probability versus magnitude. Risk averse participants should respond late in CEV and early in CEVR. The first two conditions can yield important information on the relative capabilities of patients to speed or slow their reaction times (RTs).

9.5.2 Time Conflict Model

The model for the TUIT begins with the updating of state values as seen in the actor-critic model. Through experience on the task (after pressing the response button at various times) participants develop an EV \( V_t \) for a given moment in time represented on the clock face. As before, the EV is updated by a reward PE:

\[
V_{t+1} = V_t + \alpha \delta_t \\
\delta_t = r_t - V_t
\]

Just like the actor-critic model, the model used for the time conflict experiment assumes that value integration is computed by a critic, and these same PE signals train the actor in the striatum. Because of differential learning mechanisms for approach (Go) and avoid (NoGo) systems, different learning rates \( a \) for positive and negative PEs were used:

\[
\text{Go}_{t+1}(s,a) = \text{Go}_t(s,a) + \alpha^{\text{Go}} \delta^{+\text{ve}}_t \\
\text{NoGo}_{t+1}(s,a) = \text{NoGo}_t(s,a) + \alpha^{\text{NoGo}} \delta^{-\text{ve}}_t
\]

where \( \alpha^{\text{Go}} \) and \( \alpha^{\text{NoGo}} \) describe the amount of speeding and slowing of RTs in response to positive and negative PEs, respectively. As mentioned above, positive PEs are expected to engage an approach-related dopamine D1-dependent speeding effect on RT, while negative PEs are expected to engage an avoidance-related D2-dependent slowing effect on RT.

In addition to the base-RL model used, several further parameters were included to better describe the observed RTs. A parameter \( K \), describing the baseline motor response tendency: participants initial RTs generally occurred within the first half of the clock face, participants then slow to explore the remaining clock space. A parameter \( \rho \), describing RT
adjustment toward RTs obtain more positive PEs. This was estimated using Bayesian updating: participants develop a representation (or prior) of the distribution of positive PEs for fast (RT < median) and slow (RT > median) responses separately over the history of the set, these representations were then updated with new incoming evidence via Bayes rule:

$$P(\theta|\delta_1...\delta_n) \propto P(\delta_1...\delta_n|\theta)P(\theta)$$

where $\theta$, represents the parameters governing the belief distribution of reward PEs for each future response. An explore parameter $\epsilon$, describing large RT fluctuations during periods of high uncertainty. The magnitude of the uncertainty was derived from the trial-by-trial updated standard deviations derived from beliefs about obtainable rewards yielded from the Bayesian model. The explore parameter was given by:

$$\text{Explore}_t(s) = \epsilon \left[ \sigma(\delta|s,a=\text{slow}) - \sigma(\delta|s,a=\text{fast}) \right]$$

where $\sigma_{\text{slow}}$ and $\sigma_{\text{fast}}$ are the standard deviations that represent outcome uncertainty after slow and fast relative responses. A parameter $\lambda$, describing the impact of the previous response’s RT on current RT. Finally, a parameter $\nu$, describing behavior that is drawn toward the RT that previously gave the largest reward, and which is at least 1 standard deviation above previous rewards. The final TUIT model was given as:

$$RT_t(s) = K + \lambda RT_{t-1}(s) - Go_t(s) + NoGo_t(s) + \rho \left[ \mu_t^{\text{slow}}(s) - \mu_t^{\text{fast}}(s) \right]$$

$$+ \nu [RT^{\text{best}} - RT^{\text{avg}}] + \text{Explore}_t(s)$$

9.5.3 Findings

As expected, controls demonstrated RT speeding during the DEV condition (Fig. 9.3A; Strauss et al., 2011). RT speeding during the DEV condition was not seen in patients, consistent with impaired Go-learning in patients (Fig. 9.3A). For the remaining conditions, there was little slowing in either group during the IEV condition, and there were no differences between patients and controls on the control conditions (CEV or CEVR). A higher burden of negative symptoms (categorized by a median split) was also associated with greater differences in the DEV condition compared to controls, with LNS patients in between (although, the correlation between scale for the assessment of negative symptoms (SANS) and DEV was not significant). This pattern of results is suggestive of a Go-learning deficit associated with the D1 subsystem in patients with negative symptoms. However, the modeling of the behavior indicated that the largest difference between patients and controls was on the
explore parameter $\epsilon$ (see Fig. 9.3B) and this parameter correlated with SANS avolition–anhedonia scores (although there were no differences using the median split). As mentioned above, this parameter reflects the degree to which exploration occurs in proportion to uncertainty about reward. With respect to Go- and NoGo-learning rates specifically, there was only weak evidence for a reduction in $\alpha^\text{Go}$ in patients (no differences in $\alpha^\text{NoGo}$) and no association with negative symptoms, suggesting that the explore parameter may be a larger contributor to patient-control differences in RL performance than alterations the “Go” system.

Given that patients showed a modest deficit in $\alpha^\text{Go}$ and DEV, but not $\alpha^\text{NoGo}$ or IEV, it would seem that patients with schizophrenia possess specific RL differences in speeding RTs for reward. Moreover, these deficits are at least partly more pronounced in patients with high number of

---

**FIGURE 9.3** Results for the temporal utility integration task experiment. (A) Subjects learn to adjust the timing of their responses over trials toward the maximal expected value. This is especially notable for the decreased expected value (DEV) condition where controls show the required speeding of reaction times (RTs) to maximize reward, while patients do not. (B) The modeling replicates faithfully the pattern of response times seen in the data across all subjects. Also shown is the difference between patients and controls in the modeled parameter $\epsilon$, representing the degree of exploration during uncertainty. This parameter was correlated with scale for the assessment of negative symptoms avolition–anhedonia. CEV, constant expected value; CEVR, constant expected value reversed; CN, control; IEV, increased expected value. Reprinted from Strauss, G.P., Frank, M.J., Waltz, J.A., Kasanova, Z., Herbener, E.S., Gold, J.M., 2011. Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. Biol. Psychiatry 69, 424–431.
negative symptoms (albeit not correlated with symptoms). This deficit is suggestive of reduced sensitivity to positive PEs potentially by impairments in the D1-driven basal-ganglia pathway. Ultimately, this deficit would lead to abnormalities in using positive feedback to guide behavior with the clinical expression of this strategy manifesting as negative symptoms. By contrast the lack of differences between groups on the IEV condition is suggestive of intact D2-driven NoGo-learning and processing of negative feedback. However, the lack of a noticeable IEV effect in controls dampens the robustness of this conclusion. The results also indicate that a reduced likelihood for exploration may be feature of schizophrenia associated with negative symptom expression. High avolition—anhedonia scores were associated with a failure to strategically adjust responses to gather more information about uncertain alternative behaviors; a function served by prefrontal cortex. It was therefore argued that patients with anhedonia may assign negative value to uncertainty, potentially limiting their capacity to explore actions that would lead to significantly better outcomes.

9.6 PAVLOVIAN BIAS—EXTENDED Q-LEARNING

9.6.1 Rationale

The TUIT and the probabilistic selection task presents evidence suggesting that negative symptoms might be due to an inability to use positive PEs, particularly for states associated with a “Go” action. However, despite seeming to be selective for active or “Go” responding, none of the above mentioned tasks has explicitly used the complete withholding of an action to indicate learning. There are significant innate “Pavlovian” biases that modulate the behavior observed during operant learning tasks, such that reward-predicting stimuli invigorate and loss-predicting stimuli inhibit active or “Go” responding (Guitart-Masip et al., 2012; Cavanagh et al., 2013). Just as in the TUIT task, action invigoration in response to positive PEs and action inhibition in response to negative PEs are linked with phasic dopamine firing and the propensity to take action. The Pavlovian bias discussed here describes a propensity for more efficient learning when requiring an active response to rewards and inhibiting a response to punishments. If we cross these associations, such that rewards are obtained by withholding action and losses are avoided by initiating an action, learning becomes less efficient requiring an external system to override the innate action bias. One region of override has been located in the prefrontal cortex, which is engaged more strongly when stimuli that possess an anti-Pavlovian bias are presented.
An investigation into the role of Pavlovian biases therefore has the potential to examine several possible competing accounts of poor RL performance in schizophrenia. For example, it is possible that reward-learning deficits could be explained by a reduction in Pavlovian biases modulating action, rather than due to degraded representations of action values. There is also the possibility that alterations in dopamine signaling in schizophrenia, inherent in the illness or related with dopamine D2 receptor blockade, might dampen Pavlovian biases even in the presence of a prefrontal cortical dysfunction that fails to override the innate Pavlovian bias. Such an effect would potentially show in a situation where dampening action invigoration in response to rewarding stimuli is beneficial. This would be able to determine to some degree the generality of reward specific learning deficits in schizophrenia. Moreover, given the above findings of intact learning to negative outcomes, we may also be able to determine whether this is specific to anti-Pavlovian conditions or whether the behavior generalizes to a more explicit requirement to withhold responding.

The Pavlovian bias task reported on below presents the participant with four stimuli. Two stimuli yield rewards at a 0.8/0.2 ratio for correct responses, two stimuli yield punishments at the same ratio for incorrect responses. Two possible options are allowed by the participant, they are either required to press a button, or withhold pressing a button within the time frame allowed. Response options and stimulus-feedbacks are crossed, to give four conditions: “Go-to-Win,” “Go-to-Avoid,” “NoGo-to-Win,” and “NoGo-to-Avoid.” The Go-to-Win and the NoGo-to-Avoid conditions are the Pavlovian consistent conditions. The Go-to-Avoid and the NoGo-to-Win are the Pavlovian inconsistent conditions, with NoGo-to-Win being the most difficult condition for subjects to learn.

9.6.2 Pavlovian Bias Model

Like the above models, we begin with the backbone of a Q-learning module keeping track of state-action pairings, with PEs updating the Q-values as a function of the learning rate $\alpha$. Additionally, differential sensitivities to rewards and punishments are modeled in the PE calculation using separate reward and punishment PEs:

$$Q_{t-1}(s,a) + \alpha \left( \rho^{Rew|Pun} r_t + Q_{t-1}(s,a) \right)$$

$$\delta_{t}^{rew} = \rho_{t}^{rew} \times r_t - Q_t(s,a)$$

$$\delta_{t}^{pun} = \rho_{t}^{pun} \times r_t - Q_t(s,a)$$

where $\rho^{Rew|Pun}$ describe the scaling applied to rewards or punishments. There is now the possibility to withhold responding as a valid “action,”
we include a Go bias $b$ that modulates the state-action value for go stimuli, such that:

$$Q_t(s_{Go}) = Q_t(s_{Go}) + b$$

reflecting the degree to which participants favor active responding over withholding a response. A final addition to the model is the inclusion of a Pavlovian bias parameter $\pi$. To include this parameter, we calculate the overall experienced value of a stimulus (similar to the actor-critic model) by:

$$V_t(s) = V_{t-1}(s) + \alpha \left( r_t^{Rew|Pun} + V_{t-1}(s) \right)$$

We then modify the state-action value of Go associated state-action pairs as a function of the stimulus value weighted by the Pavlovian bias $\pi$:

$$Q_t(s_{Go}) = Q_t(s_{Go}) + b + \pi V(s)$$

thus modeling the degree of Pavlovian bias by modifying state-action values as a function of their overall history of reward or punishment.

### 9.6.3 Findings

Fig. 9.4 presents the behavioral results comparing patients and controls across trials. Within each of the four conditions, patients performed worse than controls on the Pavlovian consistent conditions “Go-to-Win” and “NoGo-to-Avoid” and actually performed similarly or better on the Pavlovian inconsistent conditions “Go-to-Avoid” and “NoGo-to-Win” Fig. 9.5A (Albrecht et al., 2016). This is a complex mix of replication and nonreplication: we were able to replicate the poor reward-based learning in patients, but this only existed on the Go condition, with opposite results in the NoGo condition; similarly, we were able to replicate a relatively intact loss-avoidance learning in the Go condition, but there was a surprising performance impairment in the NoGo condition. Overall, this resulted in the result that patients had less Pavlovian bias compared with controls (Fig. 9.5B). However, the modeling indicated only weak differences between the two groups on the fitted Pavlovian bias parameter. Instead, the most robust differences between the two groups were reductions in the Go bias and Punishment sensitivity parameters in patients. It seems this combined reduction may be responsible for the observed Pavlovian bias reduction in the patient group as a whole, a reduction in Go bias gives a disadvantage to patients during Go-to-Win but an advantage during NoGo-to-Win conditions, with the reduced punishment sensitivity necessary to model the poor performance in the patients on the NoGo-to-Avoid condition.
There were no correlations found between negative symptoms and reward sensitivity or punishment sensitivity. Only the learning rate parameter was significantly associated with negative symptoms, indicating a general impairment in updating EV, potentially similar to the $\alpha_{\text{Go}}$ findings from the TUIT task. Moreover, behavioral performance on the Go-to-Avoid condition was correlated negatively with negative symptoms, contrary to our previous results indicating that negative symptoms are associated with intact learning from punishments. There were, however, interesting effects in a subgroup of patients taking the antipsychotic clozapine. Notably, all of the above group differences were amplified in the patients on clozapine. Moreover, analysis of the modeling output managed to identify a significant reduction of the Pavlovian bias parameter in those taking clozapine, suggesting a significant impact of the type of antipsychotic in RL. With no definitive unique mechanism of
action for clozapine yet found, it is unknown exactly which target or targets may be responsible for the findings in this study. Overall, we suggested that the pattern of results seen in schizophrenia on this Pavlovian bias task is most likely due to disrupted communication between the striatum and frontal cortex and between striatal dopamine driven linkage of feedback with behavior that has been previously seen in RL tasks (Schlagenhauf et al., 2009; Quideé et al., 2013; Yoon et al., 2013). This is as opposed to enhanced functioning of the inferior frontal gyrus, one region that is likely responsible for overriding innate action biases in response to a history of feedback.
9.7 DIRECT ADDITION OF WORKING MEMORY TO REINFORCEMENT LEARNING MODELS

9.7.1 Rationale

The tasks presented above often use the repeated presentation of two to four stimuli interspersed amongst one other. This presents a moderate demand on WM systems that likely play a significant role during the kinds of RL tasks we have described above, particularly in patients with schizophrenia who frequently show deficits in WM compared with controls. This presents a potentially important confound when attempting to determine the relative contributions of different neural systems at play in schizophrenia and whether they play a role in negative symptom formation. To more explicitly assess the contribution of WM to RL we parametrically manipulated WM load. We have done this by using multiple sets of stimuli, each set containing between one and six stimuli (Collins and Frank, 2012). For example, low WM load Set A may possess two stimuli requiring an appropriate response (e.g., by pressing button 1 or button 2), while high WM load Set B possesses five interspersed stimuli. The response accuracy for Set B naturally shows significant learning delays compared to Set A, and takes many more stimuli to be presented before plateau performance is reached.

9.7.2 Reinforcement Learning and Working Memory Model

To model this data, we can again use a mixture of two models: a classic RL module and a fast learning but capacity-limited WM module. The RL module uses many of the standard Q-learning features described above, with Action-State pairings $Q(s, a)$ derived from PEs modulated by an estimated learning rate $\alpha$, and softmax temperature $\beta$. The fast-acting WM module again uses the same Q-learning model as a base, with some adjustments. The learning rate $\alpha$, instead of being estimated, is set to a value of 1 so that it can learn a state-action relationship based on a single trial. However, this ability to learn is limited by WM capacity. To model this limitation, the probability of a stimulus being in WM is given as:

$$p_{WM} = \rho \times \min \left( 1, \frac{K}{n_s} \right)$$

where $K$ is an estimated parameter for WM capacity, $n_s$ is the set size or number of unique stimuli in the block, and $\rho$ is an estimated parameter (between 0 and 1) indicating reliance on the WM system versus the RL system. Two additional parameters are also estimated in the final model, a forgetting/decay parameter $\varphi$ for both the RL and WM modules and a
perseveration parameter to account for tendencies to repeat actions despite negative feedback.

Similar to the Q-learning/Actor-Critic hybrid described above, the RL and WM modules are integrated into one model using a mixture function. Each module generates a policy for action, one for the straight RL module and one for the WM module. The final decision policy for the combined model is a mixture function given as:

$$p_{\text{RLWM}} = p_{\text{WM}} + \frac{1}{C_0} p_{\text{WM}} p_{\text{RL}}$$

where $p_{\text{RL}}$ and $p_{\text{WM}}$ are the policies from the softmax function for the RL module and WM modules, respectively. The combined model therefore captures: (1) rapid and accurate encoding of information when a low amount of information is to be stored, and (2) the decrease in the likelihood of storing or maintaining items when more information is presented or when distractors are presented during the maintenance period.

### 9.7.3 Findings

Learning was much faster for low set sizes compared with high set sizes, as could be expected given an influence of WM on RL performance (Collins et al., 2014). Differences between the two groups were particularly notable for the time taken to reach plateau/speed of learning across set sizes. Controls reached plateau performance very early at set sizes two and three, generally within the first three exposures of each stimulus. Controls also managed to reach an almost asymptotic performance for all high set size blocks after seven exposures. By contrast, patients learned much slower, taking between four and six exposures at the low set sizes to reach plateau. The impaired performance also continued during the higher set size blocks, failing to reach plateau accuracy after seven exposures. Notably, patients appeared to learn slower at all set sizes and were more impaired relative to controls as set size was increased suggesting a substantial influence of WM deficits in patient performance.

The modeling results confirmed the role of WM deficits on patient RL performance. The fitted WM parameters of capacity, forgetting, and reliance were all found to be lower in patients compared with controls, indicating both general impairments in WM and specific impairments in using a deficient WM for RL performance. This was especially notable given the presence of a seemingly intact RL module and intact RL-associated parameters in patients. Intact RL processes suggest relatively normal functioning of the basal ganglia system in medicated patients, where slow, incremental is learning thought to be driven by dopaminergic PE signaling.
There were no correlations found between modeling parameters and negative symptoms, despite the number of computational parameters implicating WM and feedback processing deficits during RL in patients. The failure to find statistically significant correlations was suggested to be because the task is more related to rule-based WM process involving the lateral prefrontal cortex rather than the mPFC/OFC. Previous associations with negative symptoms are hypothesized to relate with tasks or constructs that are involved in representing specific reward values for each stimulus/action, functions attributed to more limbic portions of PFC, e.g., OFC (Gold et al., 2012). While this study did not show a specific negative symptom association, it does highlight alternative mechanisms for RL performance deficits in schizophrenia. Notably, these results suggest that, again, patients are less able to use feedback to guide decision-making due to a core deficit in WM appearing as an RL deficit on RL tasks. Interestingly, in a replication and extension of this study (Collins et al., submitted), patients again showed significant WM deficits but showed intact performance during a transfer phase that followed the above procedure. This transfer phase potentially reflects a more pure RL aspect of learning to probabilistic RL values.

9.8 SUMMARY

The above studies highlight the utility of computational modeling in the understanding of decision-making deficits in schizophrenia generally and the computational processes carried out by the brain that may be more specifically associated with negative symptoms. One common theme throughout these studies was a behavioral deficit in learning from rewards relative to punishments in patients with a greater negative symptom burden. This was shown by reduced ability to track value in the TUIT during the DEV condition, impaired learning and transfer performance on the probabilistic stimulus-selection task for high probability rewarding stimuli, and impaired performance on the Pavlovian bias experiment for the Go-to-Win condition. Moreover, there appeared to be intact processing of punishment stimuli for each of these tasks, as long as there was a response requirement necessitating an active response.

However, the computational modeling of the behavior identified an alternative explanation for this pattern of results. Namely, that higher order processes were probably more likely to be responsible for the associations between behavior and negative symptoms. This can be seen in the mixing parameter from the probabilistic stimulus-selection task and the explore parameter in the TUIT both being associated with negative symptom expression. These two parameters can be contrasted with the other parameters that would be more likely to reflect a differential sensitivity to rewards and punishments but were not associated with
negative symptoms, including the $\alpha^{Go}$ parameter from the TUIT and the punishment sensitivity parameter $\rho^{Pun}$ from the Pavlovian bias experiment. Therefore, from the modeling evidence presented here, we would assume that negative symptoms are a failure in two processes: the ability to adequately engage the OFC over the basal ganglia during RL (mixing parameter) and the failure to adjust strategic responding to gather information during uncertainty (explore parameter). Both these parameters are putatively served by frontal cortical regions (OFC, dorsal anterior cingulate cortex) and represent higher order cognitive functions, rather than more basic responding to rewards and punishments. Indeed, a higher order explanation of negative symptoms appears to match the updated perceptions of negative symptoms mentioned in the introduction, i.e., in the moment hedonic processes seem to be intact, but assessment of future expected enjoyment/utility are dampened. Furthermore, such deficits in higher order processes might also be expected to alter cost-benefit decision-making in people with schizophrenia. The ability to represent prospective reward value is critical to the decision to expend effort to obtain a reward. There is replicated evidence that cost-benefit decision-making is altered in schizophrenia (Fervaha et al., 2013; Gold et al., 2013; Barch et al., 2014), often as a function of negative symptom severity (Gold et al., 2013; Barch et al., 2014). Such a model of symptoms provides a meeting of clinical experience with computational psychiatry.

Separate to the issue of negative symptoms, the reinforcement learning and working memory study highlighted a generic deficit in WM that seems to play a dominant role in RL performance in schizophrenia. This study provided a more direct merge between behavior and computational modeling: performance in patients was worse during conditions that placed increasing load on WM and computational modeling identified a consistent deficit in WM parameters. Furthermore, the RL module seemed to be intact in patients, again highlighting that patient-control differences are likely reflected by differences in higher order processes. Neurophysiological evidence for the notion of intact RL with deficit higher order functioning is emerging. A number of event-related potential (ERP) studies have shown intact feedback-related negativity (FRN) in patients (Morris et al., 2008, 2011; Koch et al., 2010), an evoked potential that is elicited in response to feedback and is proportional to the size of the PE (Holroyd and Coles, 2002). Indeed, during the Pavlovian bias experiment (Albrecht et al., 2016), we showed that the FRN tracked with PE in patients during the early stages of the ERP (i.e., during the FRN), but later tracking of the ERP with PE (around the P300 time region) associated with more complex processing of feedback (Fischer and Ullsperger, 2013; Ullsperger et al., 2014) showed marked differences from controls.

It is uncertain yet whether any of these parameters/behaviors are specific to the generation of negative symptoms in schizophrenia or
whether they may also cross diagnostic boundaries. Indeed, even within
the scope of schizophrenia and negative symptoms, the idea that there
might be one common underlying mechanism for the emergence of
negative symptoms seems implausible. This seems especially so given the
highly heterogenous nature of schizophrenia (Tandon et al., 2013; Arnedo
et al., 2014; Schizophrenia Working Group of the Psychiatric Genomics
Consortium, 2014; Jablensky, 2015). Nevertheless, it is plausible that the
mechanistic understandings of negative symptoms in schizophrenia ob-
tained from computational modeling could apply to syndromes such as
derpression. In a meta-analysis summarizing the results of computational
modeling during RL in depression and anhedonia, Huys et al. (2013) find
evidence for a reduction in reward sensitivity over other modeled pa-
rameters. This suggests a substantially different origin of symptom for-
formation between the two clinical groups. In depression, deficits may occur
at lower order levels related more proximally with reward processing,
while in schizophrenia, deficits seem to occur at higher order levels
implicating more strategic deficits.

9.9 CONCLUSION

Our computational modeling work in schizophrenia has identified
novel possible mechanisms for the differences seen in RL performance
that would not otherwise be explicitly identifiable given a simple analysis
of behavior. Although there has been a start at linking RL modeling with
neurophysiology in schizophrenia, further investigations integrating
multiple imaging modalities with behavior will likely refine our un-
derstandings of the affected processes responsible for the formation of
negative symptoms. This offers a powerful approach for identifying
mechanistic pathways, where our current evidence suggests that frontal
cortical deficits are the most likely originator of negative symptom
development. Notably, this appears to be to the exclusion of deficits in
more basic reward processing function.

References

Pavlovian bias in schizophrenia: enhanced effects in clozapine-administered patients.
Arendo, J., Švrakic, D.M., del Val, C., Romero-Zaliz, R., Hernández-Cuervo, H., Fanous, A.H.,
et al., 2014. Uncovering the hidden risk architecture of the schizophrenias: confirmation in
Barch, D.M., Treadway, M.T., Schoen, N., 2014. Effort, anhedonia, and function in schizo-
phrenia: reduced effort allocation predicts amotivation and functional impairment.
REFERENCES


III. CHARACTERIZING COMPLEX PSYCHIATRIC SYMPTOMS