ORIGINAL ARTICLE

Negative Symptoms and the Failure to Represent the Expected Reward Value of Actions

Behavioral and Computational Modeling Evidence

James M. Gold, PhD; James A. Waltz, PhD; Tatyana M. Matveeva, BA; Zuzana Kasanova, BA; Gregory P. Strauss, PhD; Ellen S. Herbener, PhD; Anne G. E. Collins, PhD; Michael J. Frank, PhD

Context: Negative symptoms are a core feature of schizophrenia, but their pathogenesis remains unclear. Negative symptoms are defined by the absence of normal function. However, there must be a productive mechanism that leads to this absence.

Objective: To test a reinforcement learning account suggesting that negative symptoms result from a failure in the representation of the expected value of rewards coupled with preserved loss-avoidance learning.

Design: Participants performed a probabilistic reinforcement learning paradigm involving stimulus pairs in which choices resulted in reward or in loss avoidance. Following training, participants indicated their valuation of the stimuli in a transfer test phase. Computational modeling was used to distinguish between alternative accounts of the data.

Setting: A tertiary care research outpatient clinic.

Patients: In total, 47 clinically stable patients with a diagnosis of schizophrenia or schizoaffective disorder and 28 healthy volunteers participated in the study. Patients were divided into a high-negative symptom group and a low-negative symptom group.

Main Outcome Measures: The number of choices leading to reward or loss avoidance, as well as performance in the transfer test phase. Quantitative fits from 3 different models were examined.

Results: Patients in the high-negative symptom group demonstrated impaired learning from rewards but intact loss-avoidance learning and failed to distinguish rewarding stimuli from loss-avoiding stimuli in the transfer test phase. Model fits revealed that patients in the high-negative symptom group were better characterized by an "actor-critic" model, learning stimulus-response associations, whereas control subjects and patients in the low-negative symptom group incorporated expected value of their actions ("Q learning") into the selection process.

Conclusions: Negative symptoms in schizophrenia are associated with a specific reinforcement learning abnormality: patients with high-negative symptoms do not represent the expected value of rewards when making decisions but learn to avoid punishments through the use of prediction errors. This computational framework offers the potential to understand negative symptoms at a mechanistic level.

Arch Gen Psychiatry. 2012;69(2):129-138

Author Affiliations: Department of Psychiatry, Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore (Drs Gold, Waltz, and Strauss and Mss Matveeva and Kasanova); Departments of Psychiatry and Psychology, University of Illinois, Chicago (Dr Herbener); and Departments of Cognitive, Linguistic & Psychological Sciences and Psychiatry and Human Behavior, Brown University, Providence, Rhode Island (Drs Collins and Frank).

In the past decade, interest in the role of deficits in reinforcement learning (RL) and reward processing for understanding the symptoms of schizophrenia has been increasing.[1-4] This work has been shaped by studies[5,6] of behaving primates showing that the pattern of dopamine cell firing seems to code reward prediction errors (PEs), with cells increasing their phasic firing rates when outcomes are better than expected (positive PEs) and briefly ceasing to fire when outcomes are worse than expected (negative PEs). It is thought that positive PE signals are broadcast to dopamine cell target areas and serve to reinforce currently active motor responses and representations that are associated with better-than-expected outcomes. In contrast, transient cessations in dopamine cell activity indicate that current actions have resulted in poorer-than-expected outcomes and should be avoided. This pattern of dopamine cell firing has been successfully modeled using RL algorithms,[7-9] and there is consistent support for the notion that phasic dopamine signals modify synaptic plasticity in corticostriatal circuits associated with action selection.[10-12]

It is well documented that antipsychotic medications achieve their effect through the blockade of dopamine receptors, supporting the inference that psychosis is linked to excessive dopamine release.[13-16] It has been proposed that excessive dopamine cell firing might "reinforce" or inappropriately increase the sa-

lience of stimuli and responses, driving aberrant learning processes that contribute to psychosis.[3,17] This hypothesis has been supported by empirical evidence indicating associations between abnormal PE signaling and the presence of psychosis[18] and the severity of delusions.[19]

Waltz et al[20] argued that the negative symptoms of schizophrenia may be understood as reflecting a different RL abnormality.[2] It was found that patients with schizophrenia show reduced learning from positive outcomes compared with control subjects but do not differ from control subjects in learning from negative outcomes.[20] This pattern of performance is most pronounced in patients with high-negative symptoms. However, it is unclear whether impairments in learning from positive outcomes reflect impaired learning from positive PEs per se or a deficit in representing the expected reward value of choices themselves. Past investigations used learning tasks in which the optimal response was associated with reward receipt that should generate a positive PE and the less optimal choice was associated with an actual loss or withholding an expected reward, both of which should generate negative PEs.[20] Therefore, these earlier studies could not distinguish between a failure to learn from positive PEs and a failure in the representation of the prospective reward values during decision making.

This distinction between positive PEs and the valuation of positive outcomes when making choices maps onto the distributed neural system that is involved with decision making. In the basal ganglia, reinforcement outcomes influence subsequent behavioral choices through synaptic plasticity in response to PEs signaled by dopamine neurons. This "slow" learning system is complemented by the contribution of the orbitofrontal cortex (OFC),[21] thought to represent the expected value of potential outcomes in working memory.[22,23] These OFC value representations are believed to be more rapidly and flexibly updated than those in the basal ganglia and provide additional "top-down" influence on decision making. Therefore, reward-based decision making involves (at least) 2 separate processes, namely, a learning mechanism that reinforces choices that have led to positive PEs in the past and a representation of the expected value of a situation-action pair.[23-25]

In computational models of reward learning and decision making, the contribution of the basal ganglia system is often formalized using an "actor-critic" framework or a "Q learning" framework.[9] In the former, a "critic" evaluates the reward values of particular states, and the "actor" selects responses as a function of learned stimulus-response weights. When outcomes differ from expectations, PEs are used to modify learning in the critic itself (to better predict reward values in the future). The critic's PEs also serve to increase and decrease stimulus-response weights in the actor. With learning, this scheme allows the actor to select actions with strong stimulus-response weights (ie, those that have produced more positive than negative PEs) without representing the expected reward values of the actions themselves. In contrast, in Q learning, instead of learning the value of particular states, the expected quality ("Q value") of each action is learned separately; actions are selected by comparing the various Q values of each candidate action and probabilistically choosing the largest one. In this case, PEs are computed with respect to the expected Q value of the selected action and are used to adjust expected action value directly. Therefore, whereas the actor in the actor-critic scheme does not consider the outcome values of competing actions, the Q-learning scheme makes these fundamental. There is compelling evidence that the OFC has a critical role in representing these kinds of value representations.[23,26-28]

Although these 2 RL algorithms are similar,[9] they make different predictions about the nature of representations used in reward-based decision making that can be highlighted with appropriate task manipulations. Herein, we examine a hybrid model in which (putatively striatal) action weights are learned as a function of PEs but are also modulated by representation of the expected outcome Q value (due to putative top-down input from the OFC).[23]

Moreover, this modeling framework provides an important means of contrasting different hypotheses about the origins of the reward learning deficits found in earlier work. Specifically, deficits in learning from rewarding outcomes could be the result of a primary failure in the ability to signal positive PEs or, alternatively, impairments in the ability to represent the positive expected value of decision outcomes.

Following work by Pessiglione et al[29] and Kim et al,[30] we implemented a task in which participants were asked to simultaneously learn 4 discriminations. In 2 pairs, the choice of the optimal stimulus is probabilistically associated with the receipt of money (a positive PE), and the choice of the nonoptimal stimulus results in no reward (ie, a zero outcome). Failure to obtain a reward in these pairs should result in a negative PE. In these pairs, as in prior work, rewards and positive PEs are conflated. In 2 other pairs, the choice of the optimal stimulus results in no loss (ie, loss avoidance, a zero outcome), whereas the choice of the nonoptimal stimulus results in overt monetary loss. In this overall design, the same "zero" outcome would result in a positive PE when it occurs in the context of potential negative outcomes[31,32] but would result in a negative PE when encountered in the context of potential rewarding outcomes. This interpretation is consistent with computational models indicating that active avoidance relies on learning from positive PEs[33-35] and that avoidance of an aversive outcome activates reward areas.[30]

After the initial acquisition phase of the task, participants completed a transfer test phase in which they chose between novel combinations of all the trained stimuli without additional feedback.[36] Critically, some pairs involved selecting between an action that had been rewarding and one that had simply avoided a loss. Both actions should have produced positive PEs during learning, leading to an increased tendency to select them. If one's choices are solely determined by the strength of association with positive PEs (as in actor-critic), the rewarding stimulus and loss-avoiding stimulus would be of equal value. Alternatively, if one was sensitive to the expected outcome of the action (eg, if action selection relies on Q values), participants should prefer the gain-producing action over the action with zero outcome.

We hypothesized that patients having schizophrenia with high-negative symptoms would show specific impairment in representing reward value. This would likely implicate OFC dysfunction in these patients.[20,37]

Table. Demographic and Clinical Characteristics and Neuropsychological Test Data for Patients and Control Subjects

Variable	HC Group (n = 28)	LNS Group (n = 22)	HNS Group (n = 25)	*P* Value
Demographic and Clinical Characteristics				
Age, mean (SD), y	41.18 (9.41)	44.77 (9.13)	41.84 (10.73)	.85
Education, mean (SD), y				
Participant	14.82 (1.87)	13.05 (1.40)	11.83 (1.86)	<.001
Maternal	13.07 (2.22)	14.11 (2.76)	12.47 (2.89)	.16
Paternal	13.48 (2.65)	14.90 (3.21)	12.39 (3.82)	.05
Sex, No.				
Male	17	14	21	.15
Female	11	8	4	
Race/ethnicity, No.				
African American	13	6	12	.23
White	14	15	11	
Other	1	1	2	
Antipsychotic medication regimen, No.				
Haloperidol or fluphenazine only	. . .	2	5	. . .
Clozapine only	. . .	6	3	. . .
Other second generation	. . .	7	5	. . .
Clozapine plus another antipsychotic	. . .	5	9	. . .
Other combination	. . .	2	3	. . .
Neuropsychological Test Data, mean (SD)				
Clinical rating score				
BPRS total	. . .	34.00 (6.05)	41.08 (8.19)	.01
BPRS positive cluster	. . .	2.24 (1.08)	2.79 (1.39)	.14
BPRS negative cluster	. . .	1.69 (0.65)	2.18 (0.79)	.03
BPRS disorganized cluster	. . .	1.31 (0.28)	1.47 (0.42)	.13
SANS total	. . .	22.91 (2.35)	36.96 (10.70)	<.001
Calgary Depression Scale score	. . .	1.96 (2.10)	2.84 (3.13)	.27
Standard neuropsychology score				
WRAT	104.50 (14.55)	95.68 (15.72)	96.08 (13.93)	.06
WTAR	105.50 (13.83)	96.45 (13.32)	97.52 (15.89)	.05
WASI	110.64 (13.65)	96.55 (14.96)	97.48 (12.34)	<.001
MATRICS battery	48.50 (12.10)	28.14 (13.82)	29.80 (11.21)	<.001

Abbreviations: BPRS, Brief Psychiatric Rating Scale; HC, healthy control; HNS, high-negative symptom; LNS, low-negative symptom; MATRICS, Measurement and Treatment Research to Improve Cognition in Schizophrenia Consensus Cognitive; SANS, Scale for the Assessment of Negative Symptoms; WASI, Wechsler Abbreviated Scale of Intelligence; WRAT, Wide Range Achievement Test 4; WTAR, Wechsler Test of Adult Reading.

METHODS

PARTICIPANTS

Forty-seven patients (45 outpatients and 2 inpatients) meeting *DSM-IV*[38] criteria for schizophrenia (n=42) or schizoaffective disorder (n=5) and 28 demographically similar volunteer healthy control (HC) subjects participated in the study. All patients had been on a stable medication regimen for at least 4 weeks at the time of testing and were considered clinically stable by treatment providers. The outpatients were recruited from the Maryland Psychiatric Research Center and from local clinics. The 2 inpatients were recruited from the Maryland Psychiatric Research Center Treatment Research Unit. All were taking antipsychotic medication (**Table**). Fifteen patients were also treated with antidepressants, 9 with mood stabilizers, 7 with anxiolytic agents, and 6 with anticholinergic medication.

The HCs were recruited from the community via random-digit dialing and word of mouth among recruited participants. They had no current Axis I or Axis II diagnoses as established by the Structured Clinical Interview for *DSM-IV*–Axis I Disorders[39] and Structured Interview for *DSM-IV* Personality,[40] reported no family history of psychosis, and were taking no psychotropic medications. All participants had no history of significant neurological injury or disease and reported no significant medical or substance use disorders. All participants provided informed consent for a protocol approved by the University of Maryland Institutional Review Board.

Patients were divided into a high-negative symptom (HNS) group and a low-negative symptom group (LNS) by a median split on the sum of the avolition and anhedonia global items on the Scale for the Assessment of Negative Symptoms.[41] These items were selected because they have been found to reflect a single factor[42,43] and are more theoretically relevant to reward learning than the "restricted affect" factor of the Scale for the Assessment of Negative Symptoms.

The 3 groups did not significantly differ in age, sex, or race/ethnicity (Table). The HC group had more years of education than both patient groups, and the LNS group completed more years of education than the HNS group. Both patient groups had moderate symptom severity, as indicated by their Brief Psychiatric Rating Scale[44] total scores. When Brief Psychiatric Rating Scale factors[45] were examined, the HNS group had greater severity on the negative cluster symptom factor, while the 2 patient groups did not differ on the positive cluster or disorganized cluster factors.

NEUROPSYCHOLOGICAL TESTING

All participants completed measures of word reading,[46,47] general intelligence,[48] and the MATRICS (Measurement and Treatment Research to Improve Cognition in Schizophrenia Consensus Cognitive) battery.[49] The HC group scored significantly

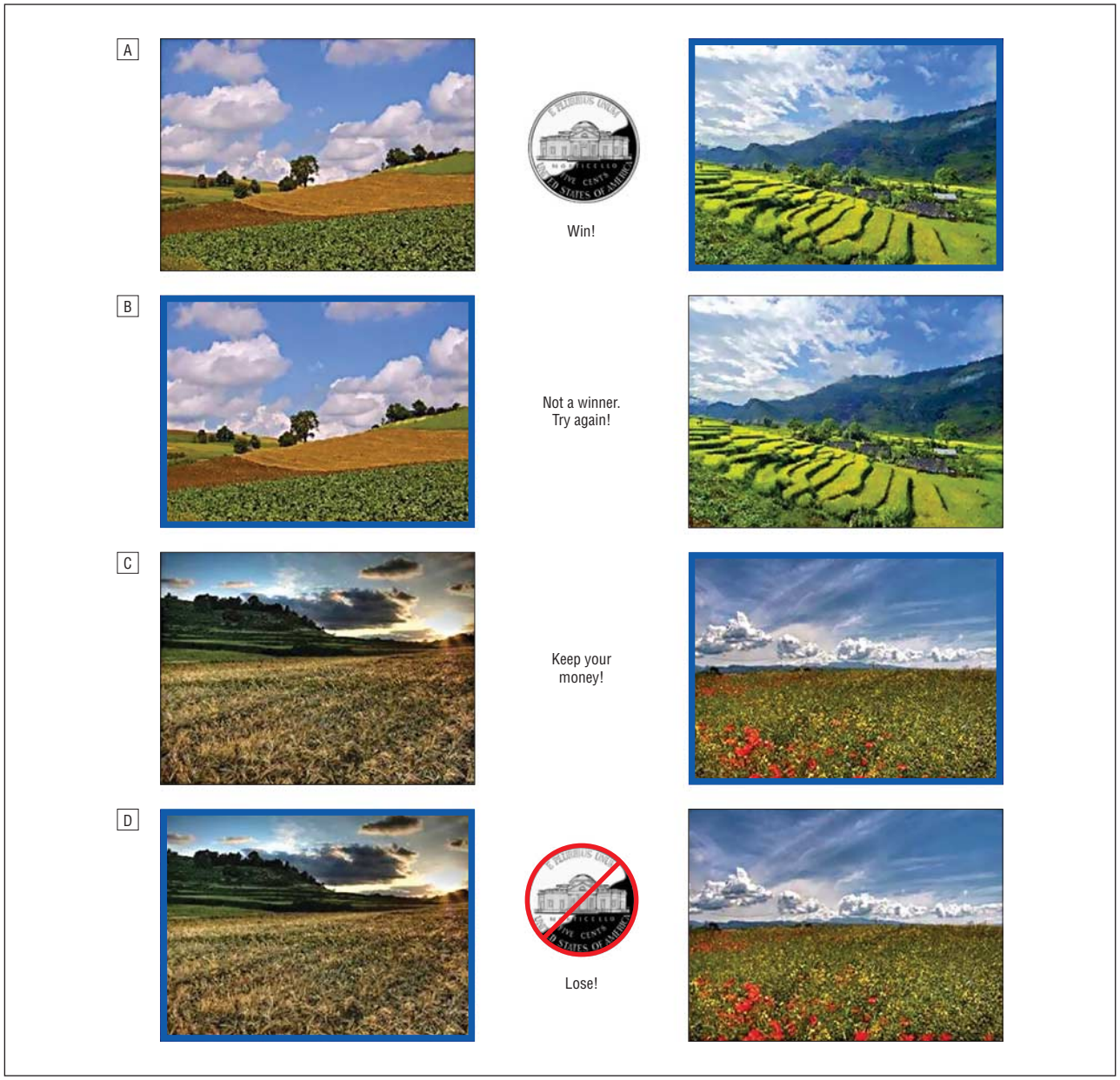


Figure 1. Example of reinforcement learning task stimuli and feedback. A, Feedback delivered after a correct choice (indicated by a blue border) in the reward trials. B, Feedback delivered following an incorrect choice. C, Feedback delivered following a correct choice in the loss-avoidance trials. D, Feedback delivered following an incorrect choice.

higher than both patient groups on all standard measures (Table). The 2 patient groups showed almost identical performance on all standard cognitive measures.

RL TASK

The learning task was administered via commercially available software (E-Prime; Psychology Software Tools) and was run on a laptop computer with a 17-in monitor. Stimuli were color images of landscapes appearing on a gray background. Participants were presented with 4 pairs of landscape items, 1 pair at a time (**Figure 1**). Two pairs involved potential gain; if the correct item was selected, participants saw an image of a nickel coupled with the word "Win!," whereas if the incorrect item was selected, they saw "Not a winner, Try again!" The correct response was reinforced on 90% of trials in one pair and on 80% of trials in the other pair. Two other pairs involved learning to avoid losses; in these pairs, selection of the correct response received the feedback "Keep your money!," whereas selection of the incorrect item resulted in the feedback "Lose!" Therefore, if the best item in the loss-avoiding pairs was selected, participants avoided a loss 90% or 80% of the time. A brief 12-trial practice session was administered to ensure task comprehension, followed by 160 learning trials with all pair types presented in a randomized order. Each pair was shown 40 times during training. To examine learning, the 160 trials were divided into 4 learning blocks of 40 trials.

Following training, the transfer test phase was presented. In these 64 trials, the original 4 training pairs were each presented 4 times, and the 24 novel pairings were each presented twice. For novel pairings, each trained item was presented with every other trained item (ie, an item that had been a 90% winner was paired with both items from the 80% gain pair, the 90% loss-avoidance pair, and the 80% loss-avoidance pair). Partici-

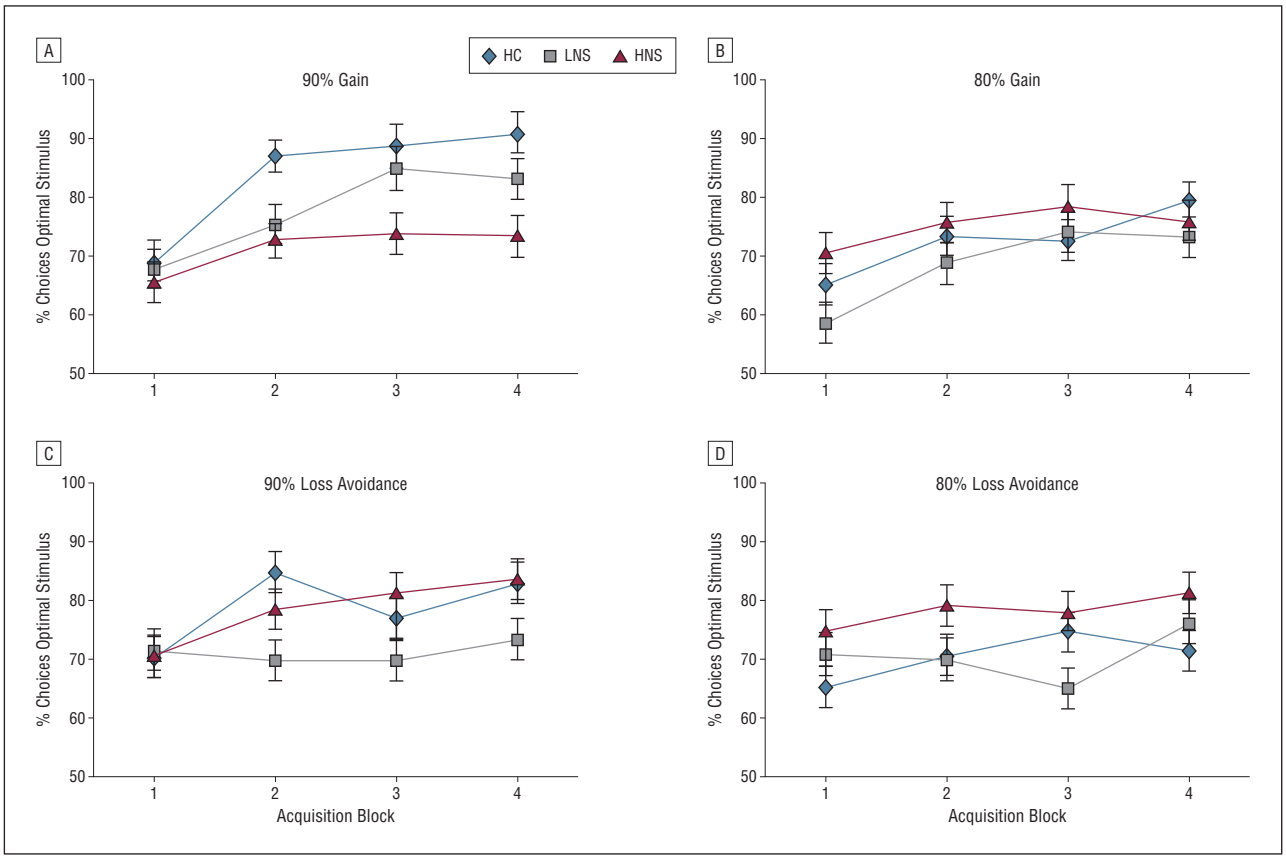


Figure 2. Differences in reinforcement learning among patients and healthy control (HC) subjects in 90% and 80% probability gain and loss-avoidance conditions. A and B, Performance in the 90% and 80% gain conditions, respectively. C and D, Performance in the 90% and 80% loss-avoidance conditions, respectively. HNS indicates high-negative symptom; LNS, low-negative symptom.

pants were instructed to pick the item in the pair that they thought was "best" based on their earlier learning. No feedback was administered during this phase.

COMPUTATIONAL MODEL

We examined the ability of the following 3 different models to fit each participant's trial-by-trial sequence of choices across training and transfer test phases: (1) a standard actor-critic architecture simulating pure basal ganglia–dependent learning, (2) a pure Q-learning model simulating action selection as a function of learned expected reward value, and (3) a hybrid model in which an actor-critic is "augmented" by a Q-learning component meant to capture the top-down influence of OFC value representations onto striatum. See the Appendix, eFigure 1, and eFigure 2 (http://www.archgenpsychiatry.com).

STATISTICAL ANALYSIS

An omnibus repeated-measures analysis of variance (ANOVA) was first conducted with a between-subject factor of group (HC, LNS, and HNS) and within-subject factors for feedback valence (gain vs loss avoidance), probability (90% and 80%), and learning block (blocks 1-4). Hyun-Feldt correction was applied if assumption of sphericity was violated; unless indicated, sphericity was not violated, and no correction was made. Significant interactions were followed by a series of ANOVA and post hoc least significant difference (LSD) test contrasts examining differences in block 4 performances. To examine the balance of learning from gain vs loss avoidance, we subtracted block 4 learning achieved from gain from that achieved from loss avoidance, testing for group differences using ANOVA, followed by within-group paired-sample *t* test. Transfer test phase performance was examined using 1-way ANOVA, followed by LSD post hoc contrasts.

RESULTS

BEHAVIORAL FINDINGS

As shown in **Figure 2**A, the HC group and the LNS group demonstrated robust learning in the 90% gain condition, with the HNS group demonstrating limited learning. In contrast, the groups performed similarly in the 80% gain condition (Figure 2B). The loss-avoidance learning blocks produced different results. Here, the HNS group matched or performed at slightly higher levels than the other 2 groups, suggesting that their learning is more effectively driven by loss avoidance than by gain seeking (Figure 2C and D).

The omnibus repeated-measures ANOVA with factors of group, feedback valence, probability, and learning block yielded main effects of probability (better performance in the 90% condition than the 80% condition; $F_{1,72}=6.08$, $P=.02$), learning block (better performance over time; $F_{3,216}=18.34$, $P<.001$), and a probability-×-group interaction (where both the HC group and the LNS group show better performance in the 90% condition than the 80% condition, while the HNS group shows similar performance with both probabilities; $F_{2,72}=3.89$, $P=.03$). In addition, there

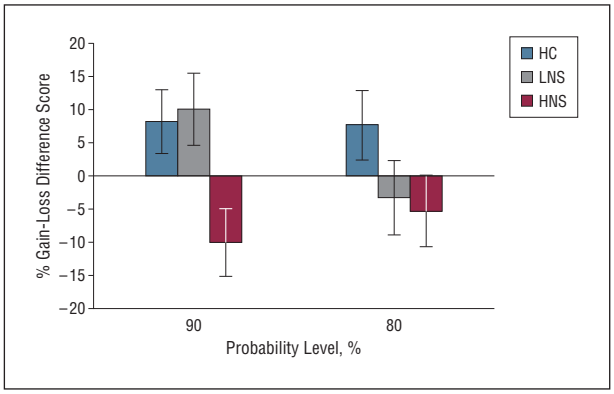


Figure 3. Performance on the gain and loss-avoidance difference score among patients and healthy control (HC) subjects. The difference score was calculated using block 4 performance. Scores above zero indicate better learning from gain than from loss avoidance, while scores below zero indicate better learning from loss avoidance than from gain. HNS indicates high-negative symptom; LNS, low-negative symptom.

was a significant feedback valence–×–learning block interaction ($F_{6,72}$=4.42, P=.005), qualified by a trend toward a feedback valence–×–learning block–×-group interaction ($F_{6,72}$=2.22, P=.06 after Hyun-Feldt correction). This last interaction suggests that the groups learned differently over time as a function of whether they were learning from rewards or from loss avoidance.

To assess whether feedback valence differentially affected final performance levels, we conducted a 2 feedback valence–×–2 probability–×–3 group repeated-measures ANOVA with block 4 performance as the dependent variable because it captured asymptotic learning levels. This analysis produced a significant main effect of probability ($F_{1,72}$=4.77, P=.03 [90% greater than 80% stimuli]) and a significant group-×–feedback valence interaction ($F_{2,72}$=4.51, P=.01) (ie, the groups learned differently as a function of feedback valence). The probability-×-group interaction fell short of significance ($F_{2,72}$=2.43, P=.10); no other effects approached significance. One-way ANOVAs examining performance for each of the 4 stimulus pairs were conducted to explore the nature of the feedback valence–×-group interaction. The only significant overall group difference was found on the 90% rewarded stimulus ($F_{2,74}$=3.83, P=.03). Post hoc LSD contrasts indicated that the HC group demonstrated significantly greater learning on this stimulus than the HNS group (P=.007); no other contrasts were significant.

To further examine feedback valence effects on learning, we computed difference scores for both the 90% and 80% conditions between end acquisition performance on gain-seeking trials and loss-avoidance trials (**Figure 3**). A positive difference score indicated better learning from gain, while a negative difference scores indicated better learning from loss avoidance. Individual 1-way ANOVAs indicated that the 3 groups differed significantly on the 90% pairs ($F_{2,72}$=4.56, P=.01). Post hoc LSD contrasts indicated significantly better learning from gain than from loss avoidance in the HC group than in the HNS group (P=.01); all other contrasts and tests of other pairs were nonsignificant.

Finally, we conducted within-group paired-sample t tests to test the comparative influence of learning achieved from gain vs loss avoidance at each probability level. There was only one statistically significant difference: the HNS group learned significantly more from the 90% loss-avoidance stimulus than from the 90% gain stimulus (P<.05).

TRANSFER TEST PHASE PERFORMANCE

Performance on 9 types of novel stimulus pairings was examined for the transfer test phase (Appendix, eFigure 1, eFigure 2, and **Figure 4**C). Pairings in which participants were confronted with the most frequently rewarded stimuli (FW in the figures) and the stimuli that most reliably avoided losses (FLA in the figures) provided the critical test of the hypothesis that the HNS group showed a specific impairment in representation of expected positive value of decision outcomes rather than learning from positive PEs. The 1-way ANOVA examining differences among the groups was significant ($F_{2,74}$=5.81, P=.005), with post hoc LSD comparisons indicating a significant difference between the HC group and the HNS group (P=.001) and an approach toward a significant difference between the LNS group and the HNS group (P=.06). As shown in Figure 4C, the HC group showed a robust preference for frequently rewarded stimuli over loss avoiders, consistent with the pattern expected if they were representing the positive expected value of the stimuli rather than relying on the number of times a stimulus has been associated with a positive PE. In contrast, the HNS group showed no preference for gain relative to loss avoiders, indicating that their preferences were based on the accumulation of positive PEs and did not take into account the value associated with those positive PEs. Although we assessed whether there were significant differences between groups in other stimulus–feedback valence comparisons, no other statistically significant differences were found.

An alternative explanation for our results is that the lack of preference for gain over loss avoidance in the HNS group might be due to difficulty in learning about rewards in general. However, as shown in Figure 4, the HNS group demonstrated a robust preference for frequently rewarded stimuli over frequently losing (FW vs FL in the figure) stimuli during the transfer test phase, with no differences observed among the 3 groups (overall $F_{2,74}$=2.06, P=.14). Furthermore, the HNS group preferred frequently rewarded stimuli over infrequently rewarded stimuli (FW vs IW in the figure). Therefore, the failure to prefer "winners" over loss avoiders cannot be explained by a failure to have learned which stimuli were associated with reward receipt.

We also examined the preference for frequent loss avoiders over infrequent winners (FLA vs IW in the figures). All 3 groups had a robust preference for the loss avoiders, despite the fact that the infrequent winner actually had a slightly positive expected value that was higher than that of loss avoiders. Therefore, all 3 groups preferred the stimulus that was more frequently associated with a positive PE over a choice that had a higher expected value but was also associated with more frequent negative PEs during learning.

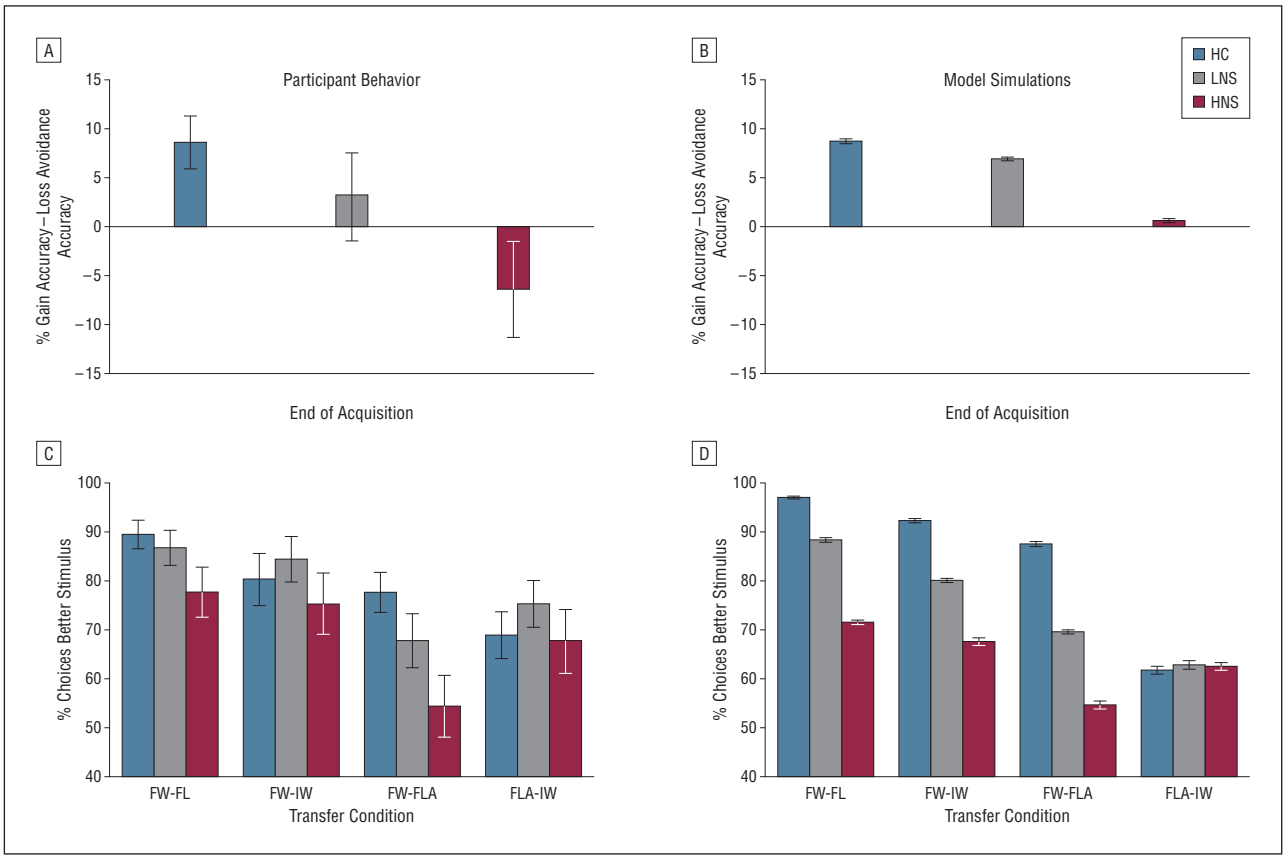


Figure 4. Observed and model simulation results for end acquisition and transfer test phase performance in patients and healthy control (HC) subjects. A and B, Observed (A) and simulated (B) end acquisition performance across groups, showing how the modeled controls had a preference for learning from gains relative to losses, an effect that is reduced in the low-negative symptom (LNS) group and absent in the high-negative symptom (HNS) group. C and D, Observed transfer test phase performance (C) and simulation results (D). Note that the simulations capture the reduced preference for frequent winners (FW) over frequent loss avoiders (FLA) in the HNS group (the only significant difference in the behavioral analyses of the transfer test phase pairs), coupled with a preserved preference for frequent winners over frequent losers (FL) and infrequent winners (IW). All groups and simulated groups show a preference for frequent loss avoiders over infrequent winners, despite having lower expected value.

EFFECTS OF ANTIPSYCHOTIC MEDICATION

We calculated haloperidol equivalents for antipsychotic medication dosage for each patient using Expert Consensus Panel guidelines.[50] There was no difference in overall antipsychotic burden between the HNS group and the LNS group ($t=0.58$, $P=.56$). Furthermore, we found no significant correlations between medication dosage and any measures of acquisition, training, or transfer test phase performance. These results suggest that antipsychotic burden is unlikely to account for our findings; however, we cannot rule out an effect of antipsychotic medication on performance that might only be observed by studying nonmedicated patients.

COMPUTATIONAL MODELING

The goal of computational modeling was to provide quantitative fits of the overall pattern of acquisition and transfer test phase data by each of 3 models (Appendix, eFigure 1, and eFigure 2). Figure 4B and D show that the best-fitting model reproduces the central features of the data in both training and transfer test phases, including better learning from gain than from loss avoidance (Figure 4B) and preference for frequent winners over frequent loss avoiders at the transfer test phase in the HC group (Figure 4D). Both of these effects are severely attenuated in the HNS group. The simple actor-critic model was insufficient for the HC group because it captured neither (1) more robust acquisition for winners vs loss avoiders (Figure 4A) nor (2) the observed robust preference for winners over loss avoiders at the transfer test phase. The pure Q-learning model could not account for the observed preference of frequent loss avoiders (FLA in the figures) compared with infrequent winners (IW in the figures) across all groups because infrequent winners have higher expected value (**Figure 5**B). The critical results are that the hybrid actor-critic–Q-learning model provided the best overall fit to the data and that the HNS group differed from the HC group and the LNS group specifically by demonstrating a reduced Q-learning component.

We tested whether the fitted parameter values from the hybrid model differed by group using ANOVA. We found a main effect of group for the mixing parameter c (Figure 5A) ($F_{2,67}=3.8$, $P=.03$), indicating a significant difference between groups in the degree to which the Q-value component influenced choices. Follow-up analyses revealed significantly lower contribution of Q values for the HNS group compared with the HC group ($t=2.77$, $P=.008$), as well as a trend in the comparison of the LNS group with the HC group ($t=1.70$, $P=.09$). As shown in Figure 5A, the HC group data were characterized by greater influence of

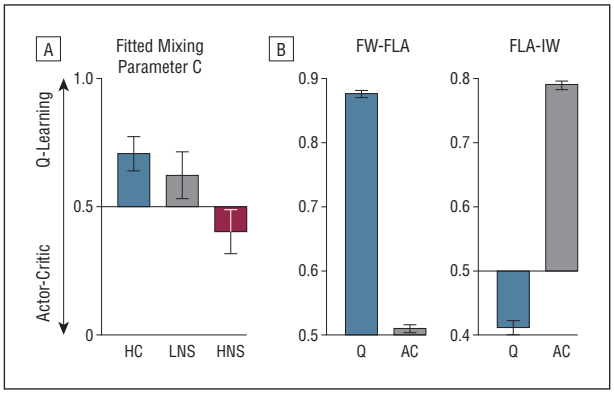


Figure 5. The relative contribution of Q learning and actor-critic learning to behavioral choices. A, Greater contribution of Q learning in healthy control (HC) subjects relative to the patient groups. Only the contrast between the HC group and the high-negative symptom (HNS) group was statistically significant. B, Predicted performance in a model of pure actor-critic (AC) or pure Q learning (Q) in the 2 diagnostic transfer test phase pairs. The Q model shows clear preference for frequent winners (FW) over frequent loss avoiders (FLA), whereas the actor-critic model does not. The 2 models show opposite preferences for frequent loss avoiders over infrequent winners (IW). One thousand model simulations were run to generate these predictions using parameters fit to the controls, but the pattern is robust to parameter changes. LNS indicates low-negative symptom.

Q learning than actor-critic learning, whereas the HNS group showed the opposite pattern.

COMMENT

These results provide insight into the origins of avolition and anhedonia in schizophrenia. First, patients with the most severe negative symptoms demonstrate deficits in learning from rewarding outcomes. This deficit is not a manifestation of a general learning impairment because the HNS group performed at levels similar to those of the HC group when learning to avoid losses. Second, in the transfer test phase, the HNS group did not show a preference for a frequently rewarded stimulus over a frequent loss avoider; that is, they were less able to take expected reward values into account during decision making; therefore, decisions were based on stimulus-response weights learned from prior PEs.

This is an RL formula for avolition: patients are better able to learn actions that lead to the avoidance of punishing outcomes than they are to learn actions that lead to positive outcomes. This pattern of data suggests that negative symptoms are not associated with reduced learning from positive PEs per se, as previously suggested, but rather with impairment in the representation of positive expected value to guide decisions. This conclusion is consistent with other data suggesting that negative symptoms are associated with deficits in reward-based tasks that depend on prefrontal or orbitofrontal cortical function.[20,37,51]

It is notable that the LNS group differed minimally from the HC group in RL behavior, with no statistically significant differences observed. Therefore, RL impairments may not be characteristic of all patients with schizophrenia but may be most evident in patients with HNS. Furthermore, the fact that the performance of the LNS group approached that of the HC group demonstrates that RL deficits are not caused by the use of antipsychotic medications: both patient groups were similarly medicated, and only the HNS group showed a deficit in learning from gain. Further study is needed in medication-free patients to address this question more definitively.

How do we account for impairment in learning from rewards with spared loss-avoidance learning in patients with HNS? Herein, the computational modeling serves to constrain our interpretation by providing a formalization of behavioral deficits grounded by a convergence of theoretical, cognitive, and neuroscientific constructs.[52] By reducing the Q-learning contribution, which is thought to reflect the top-down influence of the OFC, we were able to closely simulate the pattern of data observed in both the training and transfer test phases in the HNS group. Insofar as the role of Q learning in the model is consistent with current evidence about OFC function,[53,54] the modeling results provide proof of principle that this type of mechanism can account for the origins of severe negative symptoms. Clearly, this is an oversimplification because many other neuromodulatory systems and anatomic areas are involved in reward learning and may be implicated in the impairments documented herein. However, the modeling results demonstrate that it is possible to account for patient behavior in our task environment with a simple RL approach. The finding that patients and the HC group differed not only within the parameters of a given model but also in the best-fitting model itself implies that caution should be applied when interpreting functional imaging or behavioral data that assume that patients and control subjects are using the same neural and cognitive strategy (ie, the same model).

Overall, our data suggest that abnormalities in the reward system of patients with HNS are more strongly due to abnormalities in the cortical (representational) part of the reward system than to the basic machinery of dopamine signaling in the basal ganglia and limbic system. The representation of goal-directed action-outcome associations has been shown to rely on prefrontal cortical function,[55] and degraded prefrontal cortical representations may explain why the HNS group showed no preference for a gain-producing stimulus over a loss-avoiding stimulus, despite the fact that one was associated with a positive outcome and another with a zero outcome. These interpretations also converge with findings suggesting that negative symptoms are associated with a reduced tendency to make strategic exploratory responses to determine whether better rewards may be available than those experienced thus far,[37] the same pattern observed in healthy individuals with the *COMT* Val/Val genotype[56] and associated with prefrontal cortical activation.[57]

Other findings from our group are consistent with the results reported herein. Waltz et al[20] reported that patients with schizophrenia showed impaired learning from frequently rewarded stimuli but showed intact avoidance of infrequently rewarded stimuli. In a reanalysis of that data set stimulated by the present findings, it was clear that patients in the HNS group drove the impaired reward learning effect (Appendix, eFigure 1, and eFigure 2). Furthermore, in functional magnetic resonance imaging, Waltz et al[57] showed intact modulation of blood

oxygenation level–dependent signal response in the striatum in response to negative PEs but showed decreased signal in response to reward receipt. In addition, Waltz et al[20] and Strauss et al[37] demonstrated impairments in learning from positive rewards and spared learning from negative outcomes using "Go" vs "NoGo" learning paradigms with different behavioral end points. The present experiment extends these findings in a critical fashion by implicating the abnormal valuation of positive outcomes in patients' blunted learning from positive PEs associated with rewards.

Most work investigating PE signaling in schizophrenia has focused on the possibility that aberrant positive PEs may underlie positive symptoms.[18,19,58] Our focus is different, and the present design is not optimal for detecting aberrant positive PEs. Therefore, our results do not contradict prior studies but rather suggest that PE-driven RL models may also offer a means of understanding negative symptoms.

The origin of negative symptoms remains a major puzzle. By definition, such symptoms are the absence of normal function. Yet, such an absence must implicate the presence of an underlying causal mechanism. Our data suggest that patients with HNS fail to represent the relative value of different rewards when making decisions, while avoiding losses and punishing outcomes. This is an RL formula for avolition, likely resulting in a narrowing of patients' behavioral repertoires and a failure to activate behavior to accomplish goals.

Submitted for Publication: February 2, 2011; final revision received July 29, 2011; accepted August 4, 2011.
Correspondence: James M. Gold, PhD, Department of Psychiatry, Maryland Psychiatric Research Center, University of Maryland School of Medicine, PO Box 21247, Baltimore, MD 21228 (jgold@mprc.umaryland.edu).
Financial Disclosure: Dr Gold receives royalty payments from the Brief Assessment of Cognition in Schizophrenia and has consulted for Merck, Pfizer, Solvay, GlaxoSmithKline, and AstraZenaca.
Funding/Support: This work was supported by grant R01 MH080066 from the National Institute of Mental Health.
Role of the Sponsors: The funding organization had no role in the design or conduct of the study; the collection, management, analysis, or interpretation of the data; or in the preparation, review, or approval of the manuscript.
Previous Presentation: This study was presented at the 13th International Congress on Schizophrenia Research; April 5, 2011; Colorado Springs, Colorado.
Online-Only Material: The Appendix, eFigure 1, and eFigure 2 are available at http://www.archgenpsychiatry.com.
Additional Contributions: Sharon August, MA; Leeka Hubzin, MA; Jacqueline Kiwanuka, MBA; and Dhivya Pahwa, BA, contributed to the study.

REFERENCES

1. Barch DM, Dowd EC. Goal representations and motivational drive in schizophrenia: the role of prefrontal-striatal interactions. *Schizophr Bull.* 2010;36(5):919-934.
2. Gold JM, Waltz JA, Prentice KJ, Morris SE, Heerey EA. Reward processing in schizophrenia: a deficit in the representation of value. *Schizophr Bull.* 2008;34(5):835-847.
3. Kapur S. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry.* 2003;160(1):13-23.
4. Heinz A, Schlagenhauf F. Dopaminergic dysfunction in schizophrenia: salience attribution revisited. *Schizophr Bull.* 2010;36(3):472-485.
5. Schultz W. Predictive reward signal of dopamine neurons. *J Neurophysiol.* 1998;80(1):1-27.
6. Schultz W. Getting formal with dopamine and reward. *Neuron.* 2002;36(2):241-263.
7. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci.* 1996;16(5):1936-1947.
8. Montague PR, Hyman SE, Cohen JD. Computational roles for dopamine in behavioural control. *Nature.* 2004;431(7010):760-767.
9. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press; 1998.
10. Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci.* 2008;8(4):429-453.
11. Maia TV. Reinforcement learning, conditioning, and the brain: successes and challenges. *Cogn Affect Behav Neurosci.* 2009;9(4):343-364.
12. Cohen MX, Frank MJ. Neurocomputational models of basal ganglia function in learning, memory and choice. *Behav Brain Res.* 2009;199(1):141-156.
13. Davis KL, Kahn RS, Ko G, Davidson M. Dopamine in schizophrenia: a review and reconceptualization. *Am J Psychiatry.* 1991;148(11):1474-1486.
14. Laruelle M, Abi-Dargham A. Dopamine as the wind of the psychotic fire: new evidence from brain imaging studies. *J Psychopharmacol.* 1999;13(4):358-371.
15. Carlsson A. Does dopamine play a role in schizophrenia? *Psychol Med.* 1977;7(4):583-597.
16. Kapur S, Mamo D. Half a century of antipsychotics and still a central role for dopamine D_2 receptors. *Prog Neuropsychopharmacol Biol Psychiatry.* 2003;27(7):1081-1090.
17. Kapur S, Mizrahi R, Li M. From dopamine to salience to psychosis—linking biology, pharmacology and phenomenology of psychosis. *Schizophr Res.* 2005;79(1):59-68.
18. Murray GK, Corlett PR, Clark L, Pessiglione M, Blackwell AD, Honey G, Jones PB, Bullmore ET, Robbins TW, Fletcher PC. Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Mol Psychiatry.* 2008;13(3):239-, 267-276.
19. Corlett PR, Murray GK, Honey GD, Aitken MR, Shanks DR, Robbins TW, Bullmore ET, Dickinson A, Fletcher PC. Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain.* 2007;130(pt 9):2387-2400.
20. Waltz JA, Frank MJ, Robinson BM, Gold JM. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biol Psychiatry.* 2007;62(7):756-764.
21. Haber SN, Kim KS, Mailly P, Calzavara R. Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J Neurosci.* 2006;26(32):8368-8376.
22. Schoenbaum G, Roesch M. Orbitofrontal cortex, associative learning, and expectancies. *Neuron.* 2005;47(5):633-636.
23. Frank MJ, Claus ED. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol Rev.* 2006;113(2):300-326.
24. van der Meer MA, Johnson A, Schmitzer-Torbert NC, Redish AD. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron.* 2010;67(1):25-32.
25. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science.* 2004;304(5669):452-454.
26. Furuyashiki T, Gallagher M. Neural encoding in the orbitofrontal cortex related to goal-directed behavior. *Ann N Y Acad Sci.* 2007;1121:193-215.
27. Plassmann H, O'Doherty JP, Rangel A. Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making. *J Neurosci.* 2010;30(32):10799-10808.
28. Roesch MR, Olson CR. Neuronal activity related to anticipated reward in frontal cortex: does it represent value or reflect motivation? *Ann N Y Acad Sci.* 2007;1121:431-446.
29. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature.* 2006;442(7106):1042-1045.
30. Kim H, Shimojo S, O'Doherty JP. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol.* 2006;4(8):e233. Accessed October 31, 2011.

31. Holroyd CB, Larsen JT, Cohen JD. Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology*. 2004; 41(2):245-253.
32. Nieuwenhuis S, Heslenfeld DJ, von Geusau NJ, Mars RB, Holroyd CB, Yeung N. Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage*. 2005;25(4):1302-1309.
33. Moutoussis M, Bentall RP, Williams J, Dayan P. A temporal difference account of avoidance learning. *Network*. 2008;19(2):137-160.
34. Maia TV. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn Behav*. 2010;38(1):50-67.
35. Beninger RJ, Mason ST, Phillips AG, Fibiger HC. The use of conditioned suppression to evaluate the nature of neuroleptic-induced avoidance deficits. *J Pharmacol Exp Ther*. 1980;213(3):623-627.
36. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*. 2004;306(5703):1940-1943.
37. Strauss GP, Frank MJ, Waltz JA, Kasanova Z, Herbener ES, Gold JM. Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. *Biol Psychiatry*. 2011;69(5):424-431.
38. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington, DC: American Psychiatric Association; 1994.
39. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for* DSM-IV–*Axis I Disorders (SCID-I)*. Washington, DC: American Psychiatric Press; 1997.
40. Pfohl B, Blum N, Zimmerman M. *Structured Interview for* DSM-IV *Personality (SID-P)*. Washington, DC: American Psychiatric Press; 1997.
41. Andreasen NC. *The Scale for the Assessment of Negative Symptoms (SANS)*. Iowa City: University of Iowa; 1984.
42. Blanchard JJ, Cohen AS. The structure of negative symptoms within schizophrenia: implications for assessment. *Schizophr Bull*. 2006;32(2):238-245.
43. Sayers SL, Curran PJ, Mueser KT. Factor structure and construct validity of the Scale for the Assessment of Negative Symptoms. *Psychol Assess*. 1996;8(3): 269-280.
44. Overall JE, Gorman DR. The Brief Psychiatric Rating Scale. *Psychol Rep*. 1962; 10:799-812.
45. McMahon RP, Kelly DL, Kreyenbuhl J, Kirkpatrick B, Love RC, Conley RR. Novel factor-based symptom scores in treatment resistant schizophrenia: implications for clinical trials. *Neuropsychopharmacology*. 2002;26(4):537-545.
46. Wilkinson GS, Robertson GJ. *Wide Range Achievement Test 4: Professional Manual*. Lutz, FL: Psychological Assessment Resources Inc; 2006.
47. Wechsler D. *Wechsler Test of Adult Reading (WTAR)*. San Antonio, TX: Psychological Corporation; 2001.
48. Wechsler D. *Wechsler Abbreviated Scale of Intelligence (WASI)*. San Antonio, TX: Psychological Corporation; 1999.
49. Nuechterlein KH, Green MF. *MATRICS Consensus Cognitive Battery Manual*. Los Angeles, CA: MATRICS Assessment Inc; 2006.
50. Expert Consensus Panel for Optimizing Pharmacologic Treatment of Psychotic Disorders. The expert consensus guideline series. Optimizing pharmacologic treatment of psychotic disorders. *J Clin Psychiatry*. 2003;64(suppl 12): 2-100.
51. Maia TV, Frank MJ. From reinforcement learning models to psychiatric and neurological disorders. *Nat Neurosci*. 2011;14(2):154-162.
52. Schoenbaum G, Esber GR. How do you (estimate you will) like them apples? integration as a defining trait of orbitofrontal function. *Curr Opin Neurobiol*. 2010; 20(2):205-211.
53. Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK. A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat Rev Neurosci*. 2009; 10(12):885-892.
54. Balleine BW, O'Doherty JP. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*. 2010;35(1):48-69.
55. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat Neurosci*. 2009;12(8):1062-1068.
56. Badre D, Doll BB, Long N, Frank MJ. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*. In press..
57. Waltz JA, Schweitzer JB, Gold JM, Kurup PK, Ross TJ, Salmeron BJ, Rose EJ, McClure SM, Stein EA. Patients with schizophrenia have a reduced neural response to both unpredictable and predictable primary reinforcers. *Neuropsychopharmacology*. 2009;34(6):1567-1577.
58. Schlagenhauf F, Sterzer P, Schmack K, Ballmaier M, Rapp M, Wrase J, Juckel G, Gallinat J, Heinz A. Reward feedback alterations in unmedicated schizophrenia patients: relevance for delusions. *Biol Psychiatry*. 2009;65(12):1032-1039.

Supplementary Online Content

Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, Collins AGE, Frank MJ. Negative symptoms in schizophrenia from a failure in the representation of the expected value of rewards: behavioral and computational modeling evidence. *Arch Gen Psychiatry*. 2012;69(2):129-138.

Appendix. Supplementary materials.

eFigure 1. Patient and control mean (SD) performance (%) on reinforcement learning conditions during the transfer phase immediate testing session.

eFigure 2. Reinforcement learning performance in HNS, LNS, and HC subjects from our previous study.

eTable 1. Transfer Test Phase Performance in Each Group

eTable 2. Measures of Fit for the 3 Different RL Models

This supplementary material has been provided by the authors to give readers additional information about their work.

The following information is included in the Supplementary Materials for this manuscript: 1) Data on patient and control performance in the transfer phase ; 2) A re-analysis of our previous probabilistic selection data (Waltz et al, 2007)[1] analyzing group performance in relation to negative symptom sub-groups, 3) Details on the computational modeling methods and results.

Re-Analysis of Probabilistic Selection Data

We also re-analyzed our previous data on reinforcement learning in schizophrenia using a probabilistic selection task by looking at negative symptom sub-groups to determine whether our previous findings are consistent with those reported in the current manuscript.[1] Negative symptom groups were determined using a median split on the sum of the SANS avolition and anhedonia items. Participants included 24 healthy controls, 16 low negative symptom patients (LNS), and 16 high negative symptom patients (HNS). One-way ANOVA indicated that the 3 groups significantly differed on Choose A performance (the most frequently rewarded stimulus), $F(2, 53) = 7.67$, $p < 0.001$; however, there were no differences among groups on Avoid B performance (the stimulus that was the least rewarding) , $F(2, 53) = 0.37$, $p = 0.69$. Post hoc LSD contrasts indicated that HNS patients chose A significantly less than LNS ($p = 0.006$) or HC ($p < 0.001$) subjects; however, there were no differences between HC and LNS ($p = 0.46$). These findings are consistent with a deficit in Go learning, but intact No Go learning , which is specific to HNS patients. Thus, the re-analysis of our previous data is consistent with our major findings of the current study.

Computational Modeling

The goal of the modeling was to provide a quantitative fit to the pattern of data observed in patients and healthy controls. As described below, we investigated both a standard Actor-Critic architecture and a Q-learning architecture. Neither taken alone could account qualitatively for both healthy control and patient data. We thus investigated a mixture model of Actor-Critic and Q-learning, which leads to better qualitative and quantitative fits for all groups and explains key features of the data, as motivated in the main paper.

Actor-Critic (Basal Ganglia)

According to this model, participants update the expected value V(t) of a state context on each trial t. Each pair of stimuli presented together was represented as a state that might be predictive of the presence of gains or losses. Values are updated as a function of prediction errors using the delta rule:

$$V(s,t+1) = V(s,t) + á_C * ä\ (t),$$

where $á_C$ is the critic learning rate defining the degree to which values are updated on a trial-by-trial basis, and $ä\ (t) = outcome(t)-V(s,t)$ is the reward prediction error showing the discrepancy between expected value V for the current state s and the actual experienced outcome.

Prediction errors in the critic are also used to adjust weights in the actor as follows:

$$w(s,a,t+1)= w(s,a,t) + á_A\ * ä(t),$$

where $w(s,a,t)$ is the stimulus-response weight for the action selected in trial t producing the prediction error $ä\ (t)$ and $á_A$ is the learning rate for the actor which reflects how rapidly its weights are updated. Both learning rates lie in $[0,1]$.

In order to prevent unbound growth of the actor weights, we normalize them by the sum of absolute values, so that they remain on a [-1,1] scale (this also allows proper mixing with Q values, which are naturally bounded, in the hybrid model described below). For example, actor weight for action 1 is normalized according to $w(s,a\_1,t)\ ?\ w(s,a\_1,t) / (|w(s,a\_1,t)| + |w(s,a\_2,t)|)$. To avoid division by a null value we initialized the weights at 0.01. This value is small enough not to bias future probabilities for choosing a stimulus.

Actions are selected according to the standard softmax logistic function:

$$P(a_1,t)= e^{(w(s,a1,t)/â)} / (e^{(w(s,a1\ ,t)/â)} + e^{(w(s,a2,t)\ W/â)}),$$

where a_1 and a_2 denote actions leading to the selection of stimulus 1 or 2 and P(a_1,t) is the probability of choosing action 1. The parameter ? is the softmax temperature and controls the stochasticity of the choice function (e.g. the degree of exploration).

In agreement with previous studies, we also allow positive and negative rewards to be weighed differently. Positive feedback at trial ***t*** was encoded as *outcome(t) = 1-d*, neutral feedback as *outcome(t) = 0* and negative feedback as *outcome(t) = -d*. Thus the free parameter d indicates full neglect of negative outcomes if *d = 0*, full neglect of positive outcomes if *d = 1*, and equal weighing of positive and negative outcomes if *d = 0.5*.

We expected this model to capture both reward learning and loss avoidance, as in prior actor-critic models of avoidance. This model chose randomly at the initial trials of learning, and adjusted the weights associated with a stimulus in the actor following feedback. Weights were increased to reflect learning from positive PEs, and decreased to reflect choices that led to worse-than-expected outcomes. Thus, the model was more likely to repeat choices that led to positive PEs (winners and loss avoiders), while learning to avoid stimuli which produced negative PEs (losers and infrequent winners). It did not, however, distinguish between choices on the basis of actual expected outcome values (gain or loss avoidance, loss or absence of reward).

Q-Learning (OFC)

The Q-Learning model learns the expected value of each action directly, as a function of the prediction error difference between the current expected value of that action and the actual outcome:

$$Q(a,t+1) = Q(a,t) + á_O*(outcome(t) – Q(a,t)),$$

where $á_O$ is the learning rate for the OFC. The Q-Value is only updated for the action selected in the current trial.

Action selection occurs according to the same softmax rule as described higher:

$$P(a_1,t)= e^{(Q(a1,t)/â)} / (e^{(Q(a1,t)/â)} + e^{(Q(a2,t) W/â)}),$$

and the same weighing of positive and negative outcomes through free parameter *d* as for actor-critic is allowed.

As shown in numerous other studies, we expected this model to capture both reward learning and loss avoidance. The model learns the expected values associated with different actions in different states and is thus able to distinguish between choices on the basis of actual outcome values.

Hybrid Actor-Critic Q-Learning Model (OFC-BG interactions)

To account for effects predicted separately by the two previously described models, we propose a hybrid BG-OFC model, in which the BG functions as an actor-critic but its actor values are influenced by top-down OFC Q values. The model includes potentially symmetrical contributions of learned values from both models in the softmax function, by replacing individual contributions of each model by the mixture value:

$$H(s,a1,t)=[(1-c)*w(s,a,t)+c* Q(a,t)]: P(a_1,t)= e^{(H(s,a1,t)/\hat{a})} / (e^{(H(s,a1,t)/\hat{a})} + e^{(H(s,a2,t)/\hat{a})}),$$

where $0 = c = 1$ is a mixing parameter that determines the degree of pure BG vs. OFC contributions. In particular, with c=*0*, the model is reduced to the actor-critic, while with *c=1,* it is reduced to Q-learning. Since both Q-values and normalized weights lie in [-1,1], *c=0.5* indicates equivalent contributions of both systems.

Different model predictions

While all three models predict general reward learning and loss avoidance effects, they each contribute to specific effects observed in the data. In particular,

- The actor-critic model cannot account for sensitivity to actual outcome values, since it only uses reward prediction errors to modify the probability of selecting an action, as opposed to learning specific state action values. On contrary, the Q-learning model predicts sensitivity to actual outcome values, and therefore predicts that subjects will choose a frequent winner over a frequent loss avoider, as seen in HC and LNS subjects.
- The Q-learning model cannot account for the observed preference of frequent loss avoiders (FLA) compared to infrequent winners (IW) across all groups, since infrequent winners have higher expected outcome. In contrast, the AC model can account for this pattern , since frequent loss avoiders lead to frequent positive prediction errors, thus stronger positive actor weights for selecting the loss-avoiding symbol, whereas infrequent winners lead to frequent negative prediction errors, thus negative weights.

Thus, the hybrid model should be able to better account for observed results in patients and healthy subjects. In particular, for a given mixing parameter *c*, the model may recapitulate the preference of healthy subjects to choose frequent winners over frequent loss avoiders (as a function of Q value influences) but still capture a preference for frequent loss avoiders over infrequent winners (due to some influence of AC). Lower *c* values are expected to be associated with the pattern seen in HNS subjects, which correspond to a purer version of AC.

Model fittings

All three models were fitted to subjects' data using the standard likelihood procedure. Specifically, for each participant, we searched for the free parameters that would maximize the likelihood of their own trial-by-trial sequence of choices in both phases of the task, using multiple random starting points.

We found that the hybrid model afforded better fit than both other models for all three groups, even after correction for number of supplementary parameters (3, 4 and 6 respectively for Q-Learning, Actor-Critic and Hybrid models). In the following table, for each group, we report mean (and standard error) pseudo-r2, as well as Akaike Information Criterion (AIC) for complexity penalization.

Results

The fact that all groups benefit from the addition of actor-critic to the Q-learning algorithm is consistent with the finding that all groups showed a preference for frequent loss-avoider trials than infrequent winners. Indeed, this could only be predicted by the AC part of the model (see above).

We performed ANOVAs on the fitted parameter values of the hybrid model, with subject group as a factor. We only found a main effect of group for the mixing parameter c ($F(2,67)=3.8$, $p=.027$; all other parameters $F<2.15$, $p>0.12$). Post-hoc analyses revealed a significantly lower c value for the HNS-SZ patients compared to HC ($t=2.77$, $p=0.008$), as well as a trend for the comparison of LNS-SZ patients to HC ($t=1.7$, $p=0.09$).

Furthermore, the HC group exhibited estimated *c* values that were greater than 0.5 (HC: $c=0.70 \pm 0.06$, $t=3.1$, $p=0.005$; SZ-LNS: $c=0.62 \pm 0.09$, NS). This indicated a greater role for Q-Learning than Actor-Critic in

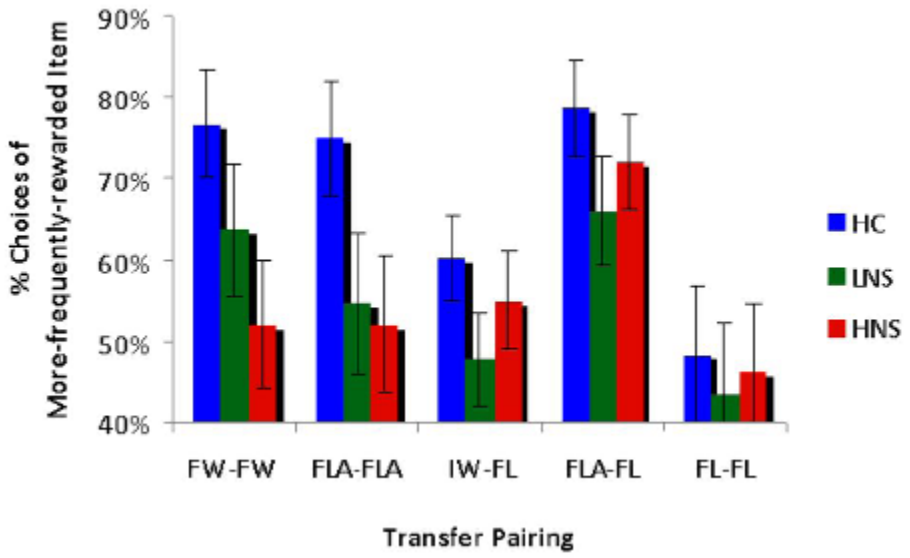
their behavior. Conversely, for the SZ-HNS group the fitted mixing parameter value ($c=0.41 \pm 0.09$) indicated a lesser role for Q-Learning than for Actor-Critic. This is consistent with the observation that those patients do not show a sensitivity to actual outcome value, contrary to HC and SZ-LNS group.

We then used the fitted parameters to simulate the hybrid model for each group, and show that the model can reproduce the key features of the observed data.

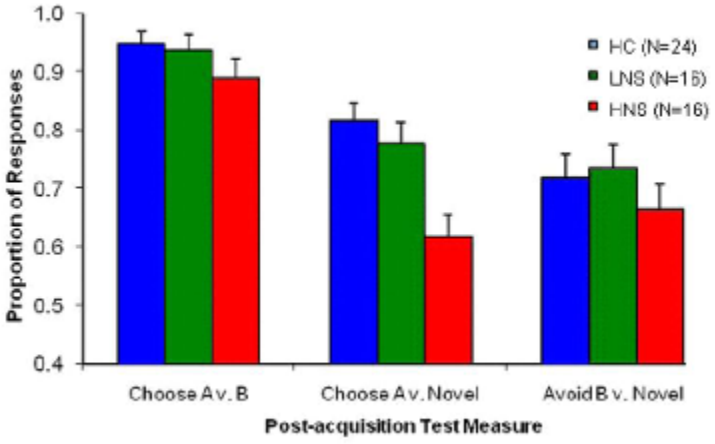
References

1. Waltz JA, Frank MJ, Robinson BM, Gold JM. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biol Psychiatry.* 2007;62(7):756-764.
2. Moutoussis M, Bentall RP, Williams J, Dayan P. A temporal difference account of avoidance learning. *Network.* 2008;19(2):137-160.
3. Maia TV. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn Behav.* 2010;38(1):50-67.
4. Frank MJ, Claus ED. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol Rev.* 2006;113:300-326.
5. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc Natl Acad Sci U S A*. 2007;104(41):16311-6. Epub 2007 Oct 3.
6. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press; 1998.

Figure Legends



eFigure 1. Patient and Control Mean (SD) Performance (%) on Reinforcement Learning Conditions During the Transfer Phase Immediate Testing Session. Note: FW = Frequent Winner; FLA = Frequent Loss Avoider; FL = Frequent Loser; IW = Infrequent Winner; AB = 90% Gain; CD = 80% Gain; EF = 90% Loss Avoidance; GH = 80% Loss Avoidance.



eFigure 2. Reinforcement Learning Performance in HNS, LNS, and HC subjects from our Previous Study.

eTable 1. Transfer Test Phase Performance in Each Group

	HC (n = 28)	**LNS (n = 22)**	**HNS (n = 25)**	***P* Value**
FW vs. FLA	78 (20)	68 (27)	53 (32)	.005
FW vs. FW	77 (35)	64 (38)	52 (39)	.061
FW vs. FL	89 (16)	86 (18)	79 (25)	.135
FW vs. IW	80 (27)	84 (23)	76 (31)	.594
IW vs. FLA	69 (26)	75 (22)	69 (31)	.660
IW vs. FL	60 (28)	48 (27)	55 (30)	.302
FL vs. FL	48 (46)	43 (42)	46 (43)	.922
FLA vs. FLA	75 (37)	55 (41)	52 (42)	.078
FLA vs. FL	79 (32)	66 (31)	72 (29)	.355
AB Pair	91 (25)	86 (23)	76 (30)	.112
CD Pair	76 (34)	77 (32)	67 (34)	.501
EF Pair	88 (19)	80 (31)	74 (33)	.213
GH Pair	78 (33)	74 (36)	76 (29)	.919

For all pairs, other than the IW vs FLA pair, the values in the table represent the percentage of trials where the participants chose the item with the highest expected value. For the IW vs FLA pair, the value shown is the percentage of choices of the FLA stimulus.

eTable 2. Measures of Fit for the 3 Different RL Models

group	Measure	Hybrid	AC	Q-Learning
HC	pseudo-r2	0.365 (0.039)	0.336 (0.037)	0.311 (0.034)\
	AIC	5528	5774	5996
SZ-LNS	pseudo-r2	0.267 (0.036)	0.2536 (0.034)	0.19876 (0.029)
	AIC	5016	5107	5479
SZ-HNS	pseudo-r2	0.260 (0.033)	0.246 (0.033)	0.190 (0.029)
	AIC	5755	5861	6293