Computational Analyses of Learning and Motivation: Lessons from Psychosis Research

James A. Waltz, PhD

Associate Professor

Maryland Psychiatric Research Center

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Where we all come from…

Partners in Crime

Jim Gold, MPRC
Univ. of Maryland

Anne Collins, UC Berkeley

Matt Albrecht,
Curtin Univ., W. Aust.

Michael Frank,
Brown Univ.

Dennis Hernaus
Univ. of Maastricht

Outline of Talk

I. Why did we start down this road?

II. How do we do what we do?

III. Modeling Probabilistic RL in a Stable Environment

IV. Modeling Probabilistic RL in an Unstable Environment

V. Modeling Directed Exploration

VI. What lessons have we learned?

VII. What do we still want to know?

What I hope you will get from my talk

- What we think the value of computational psychiatry is
- How we go about trying to address our problems of interest
- What issues we need to consider every time we apply computational approaches to a problem

“Branches” of Computational Psychiatry

- Machine learning approaches to clustering and prediction

- Neural network/“Connectionist” models of information processing

- Computational models of learning and inference
  - Rescorla-Wagner-type Reinforcement Learning models
  - Hierarchical Gaussian Filter models
  - Markov chain Monte Carlo methods
  - Drift Diffusion models

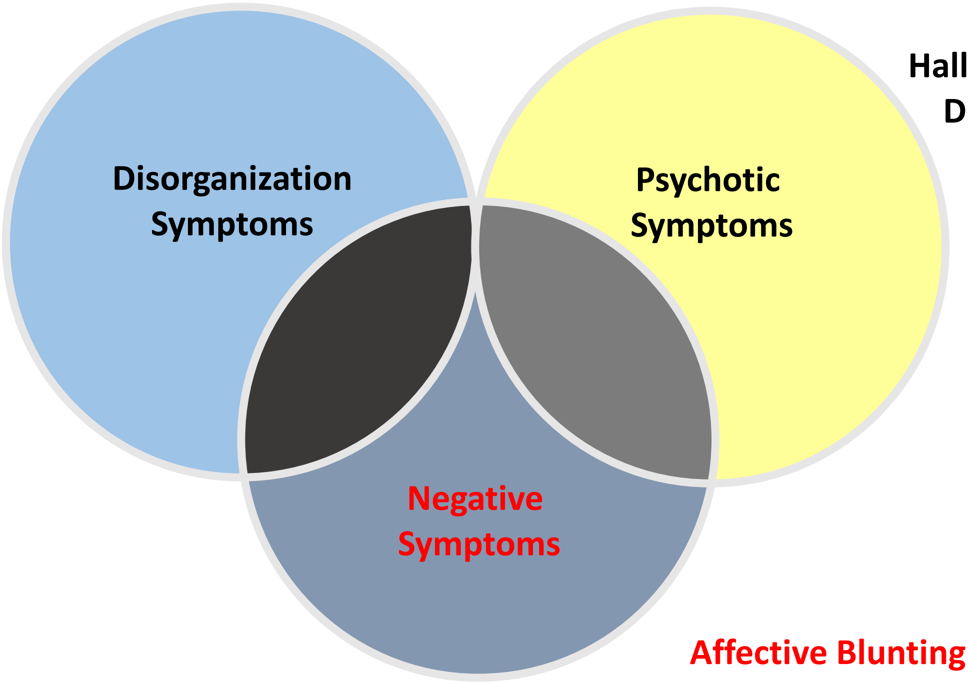
I. Why did we start down this road?

Why take this approach to studying schizophrenia?

Why care about computational accounts of learning and motivation?

Schizophrenia is a heterogenous syndrome, with multiple symptom domains

Domains of psychopathology in the schizophrenia syndrome



Hallucinations
Delusions

Affective Blunting
Anhedonia
Avolition

- SZ also generally accompanied by cognitive deficits:
  - Attention
  - Memory
  - Processing Speed
  - Executive Functions

Very little is true of MOST people with schizophrenia

Motivation to Focus on Negative Symptoms

- Negative symptoms have a high social and financial cost
  - Poor functional outcome (social, occupational; Lysaker, 2004; Norman, 2000)
  - Poor quality of life (Katschnig, 2000; Orsel, 2004)
  - Low rate of recovery (Strauss, 2011)
  - Toll on families

- No drug has received FDA approval for an indication of negative symptoms
  - 2nd generation antipsychotics not proven effective
  - Many attempts with experimental compounds, none proven consistently effective

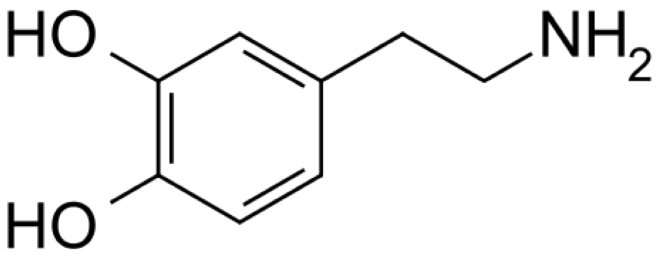
- Our understanding of *mechanisms* of negative symptoms has historically been poor, lacking in actionable targets

Particular Focus on the experiential/motivational negative symptoms of SZ

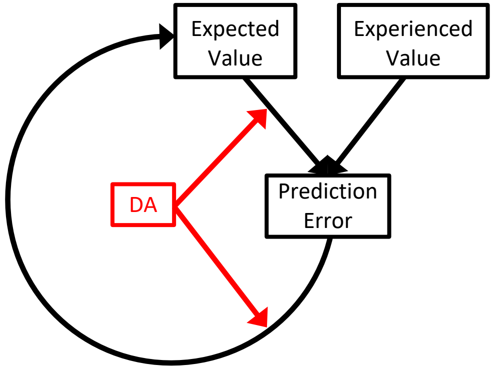
- By these, we usually mean:
  - Anhedonia: the reduced experience or anticipation of pleasure
  - **Avolition/Apathy**: reduced tendency to engage in motivated or goal-directed behavior
- Distinct from “expressive negative symptoms”, like alogia, reduced gestures, and blunted facial expressions
- Historically, the construct of anhedonia has been poorly-specified by clinical rating scales of negative symptoms (that may be changing)
  - Consummatory aspects of pleasure (“liking”) can be distinguished from anticipatory aspects of pleasure (“wanting”)
- The relationship between anhedonia and avolition is also poorly-specified, though anhedonia and avolition consistently load together as one factor in factor analyses of negative symptoms (e.g., Blanchard & Cohen, 2006 SZ Bull)
  - Does the reduced experience/anticipation of pleasure *drive* motivational deficits?

Motivating Hypotheses

- Somehow, someway, schizophrenia is a disease of dopamine systems
  - There are dopamine hypotheses of schizophrenia and psychosis
  - All antipsychotic drugs block D2 dopamine receptors and their potency as antipsychotic drugs is directly tied to their affinity for D2 receptors



- What do we know about the functional roles of dopamine pathways?
  - They appear to be involved in the signaling of reward prediction errors (RPEs)
  - They appear to signal incentive salience



- People with schizophrenia have a hedonic deficit, but the hedonic deficit is not primarily one of experience

Hedonic Experience (“Liking”) vs. Incentive Salience (“Wanting”)

Liking
(Experiencing the Reward)

The Pudding Face

vs.

Wanting
(The Cue as Motivating)

JELL-O SINGLES
DO-IT-YOURSELF DELICIOUS.
JELL-O SINGLES Chocolate
got milk?
Introducing sugar free JELL-O Pudding Singles. An 80 calorie snack you make with fat free milk. Also available in Regular.
EVERY DIET NEEDS A LITTLE WIGGLE ROOM.
Visit jello.com

Want. Pudding.

- Do people get the pudding face?
- Do people who get the pudding face when they eat it, WANT the pudding, when they are reminded of it?
- If not, it would suggest a fundamental in the ability to translate experienced reinforcement into the expectation of a reward (or approach behavior, at least)
- There is evidence that this is actually characteristic of people with schizophrenia

Schizophrenia Bulletin vol. 36 no. 1 pp. 143–150, 2010
doi:10.1093/schbul/sbn061
Advance Access publication on June 17, 2008

Emotional Experience in Patients With Schizophrenia Revisited: Meta-analysis of Laboratory Studies

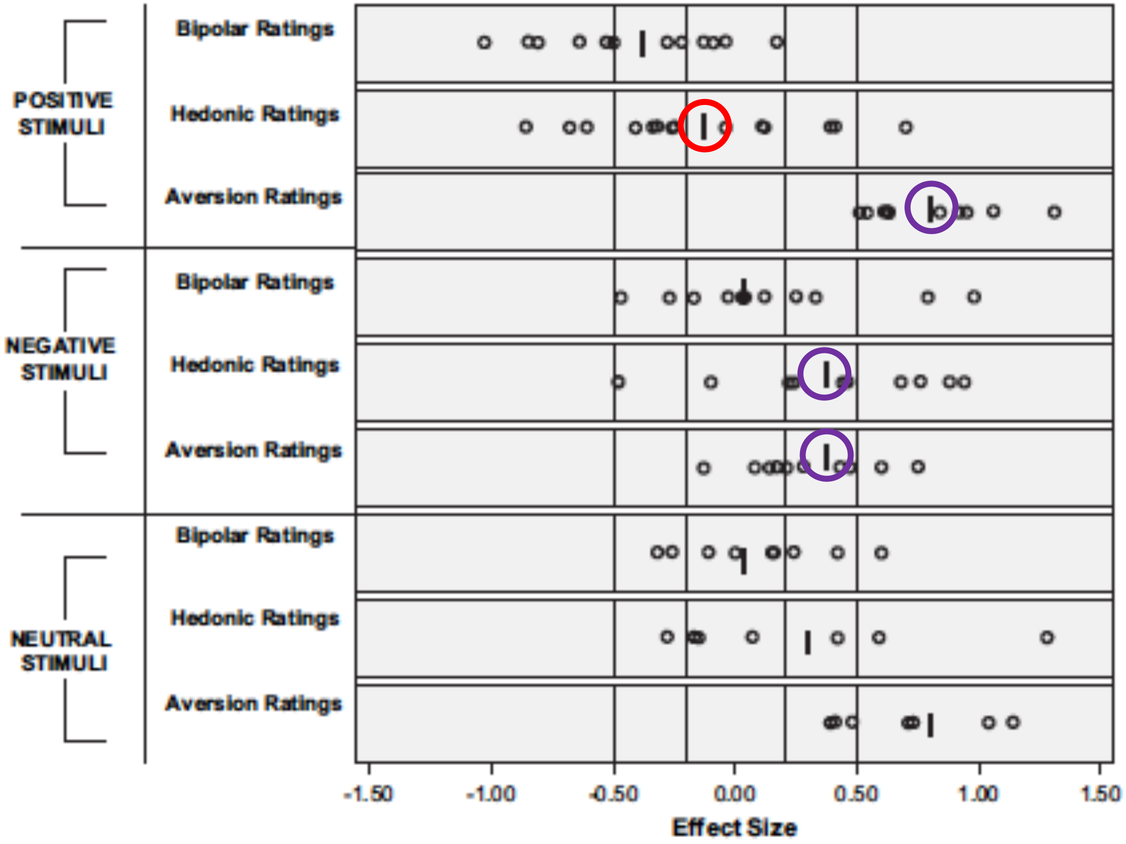


Fig. 1. Patients vs Controls: Effect Sizes Computed for Unipolar Hedonic, Unipolar Aversive, and Bipolar Emotion Ratings from the Positive, Negative, and Neutral Emotion Induction Conditions. Positive effect size values from hedonic and bipolar ratings reflect patients reporting more euphoria than controls following stimulus presentation. Positive effect size values from aversive ratings reflect patients reporting more dysphoria than controls following stimulus presentation. Dotted lines denote small (−.20 and .20) and medium (−.50 and .50) effect sizes. Dark solid line reflects weighted means.

Available online at www.sciencedirect.com

ScienceDirect

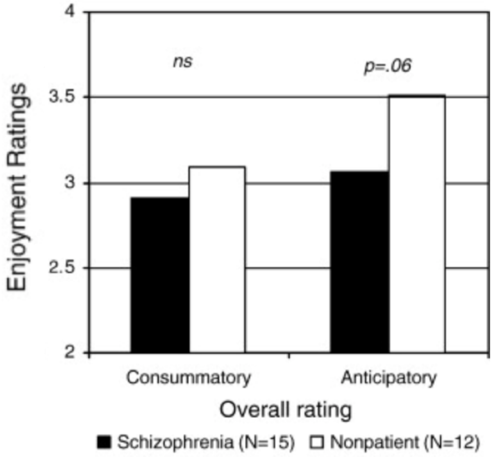
Schizophrenia Research 93 (2007) 253 – 260

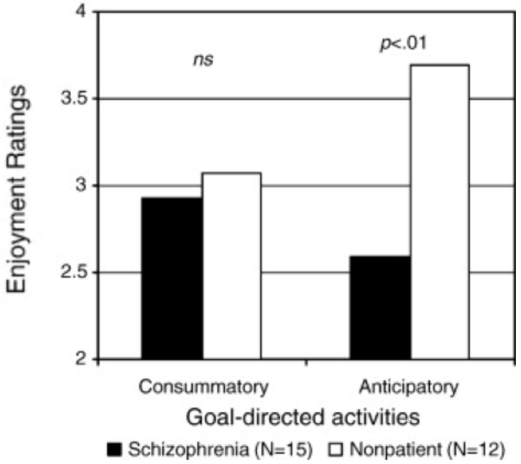
SCHIZOPHRENIA
RESEARCH

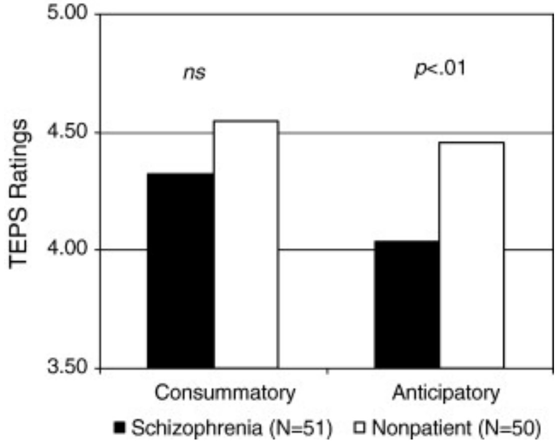
www.elsevier.com/locate/schres

Anhedonia in schizophrenia: Distinctions between anticipatory and consummatory pleasure ☆

David E. Gard [a,*], Ann M. Kring [b], Marja Germans Gard [b],
William P. Horan [c], Michael F. Green [c]







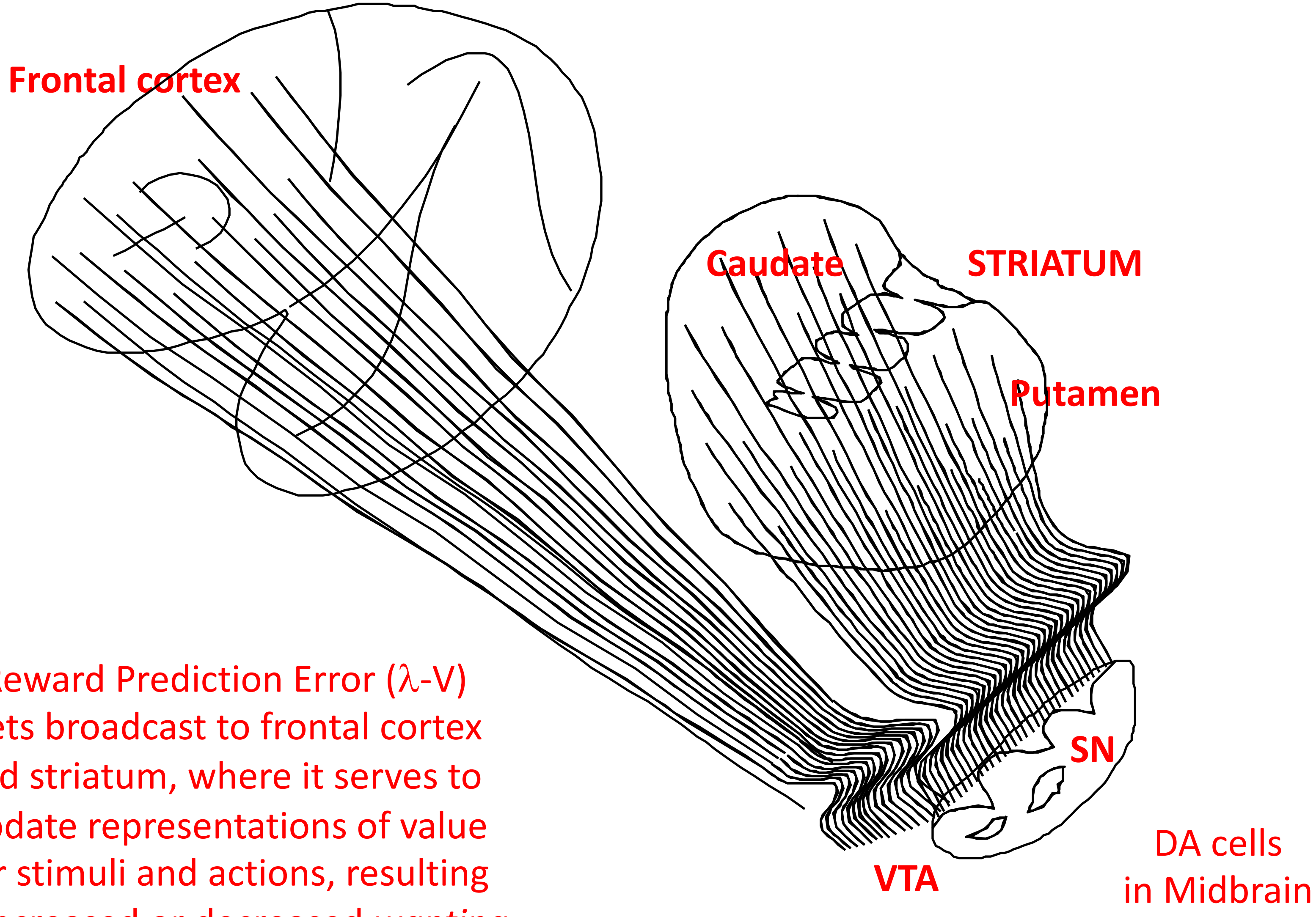
Why would SZ patients not pursue rewards that they claim to find pleasurable?

Do they not learn to “want” what they “like”?

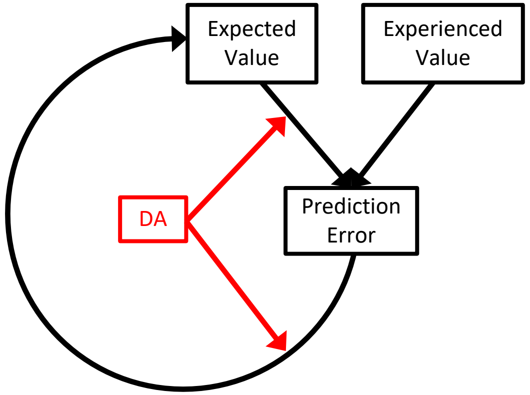
How does one learn to “want” what he “likes”?

This is the process of reinforcement learning (RL)
and there are a multitude of ways in which it can go wrong

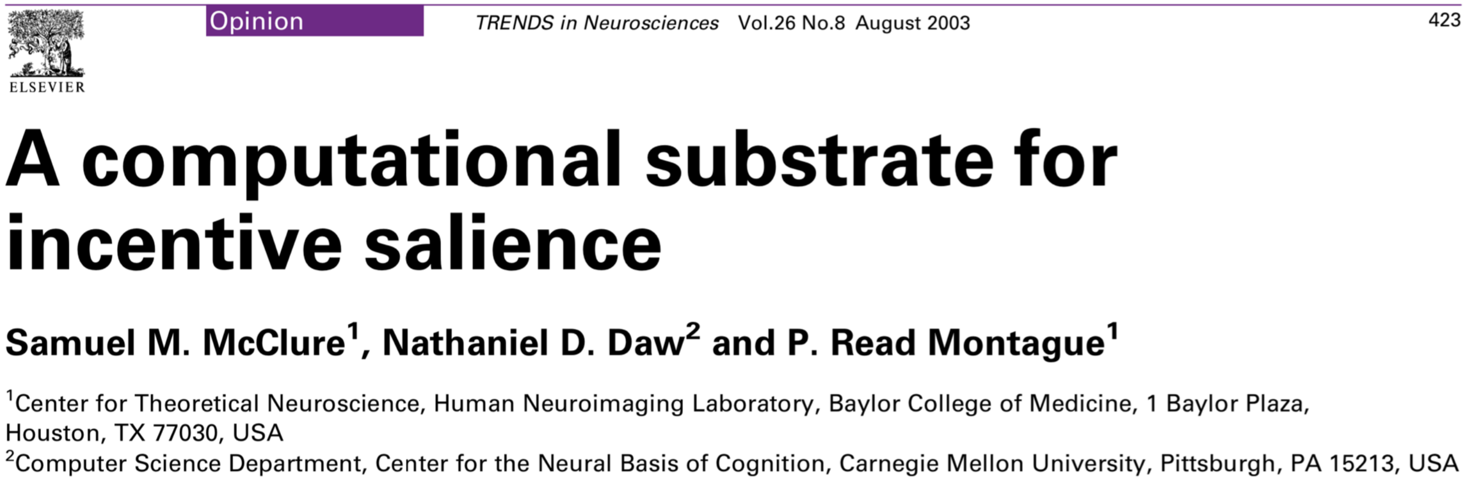
DA Neurons and the Signaling of Reward Prediction Errors (RPEs)



Reward Prediction Error (λ-V)
gets broadcast to frontal cortex
and striatum, where it serves to
update representations of value
for stimuli and actions, resulting
in increased or decreased *wanting*



Functional Roles of Dopamine:
Signaling of TDEs and Incentive Salience



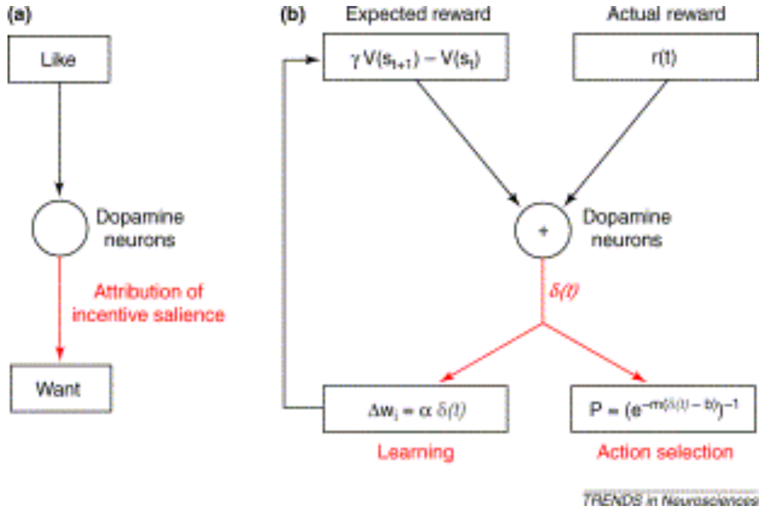
Opinion TRENDS in Neurosciences Vol.26 No.8 August 2003 423

ELSEVIER

A computational substrate for incentive salience

Samuel M. McClure[1], Nathaniel D. Daw[2] and P. Read Montague[1]

[1]Center for Theoretical Neuroscience, Human Neuroimaging Laboratory, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA
[2]Computer Science Department, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA



(a)
Like
Dopamine neurons
Attribution of incentive salience
Want
(b)
Expected reward
Actual reward
$\gamma V(s_{t+1}) - V(s_t)$
$r(t)$
+
Dopamine neurons
$\delta(t)$
$\Delta w_i = \alpha\,\delta(t)$
$P = (e^{-m(\delta(t) - b)})^{-1}$
Learning
Action selection

TRENDS in Neurosciences

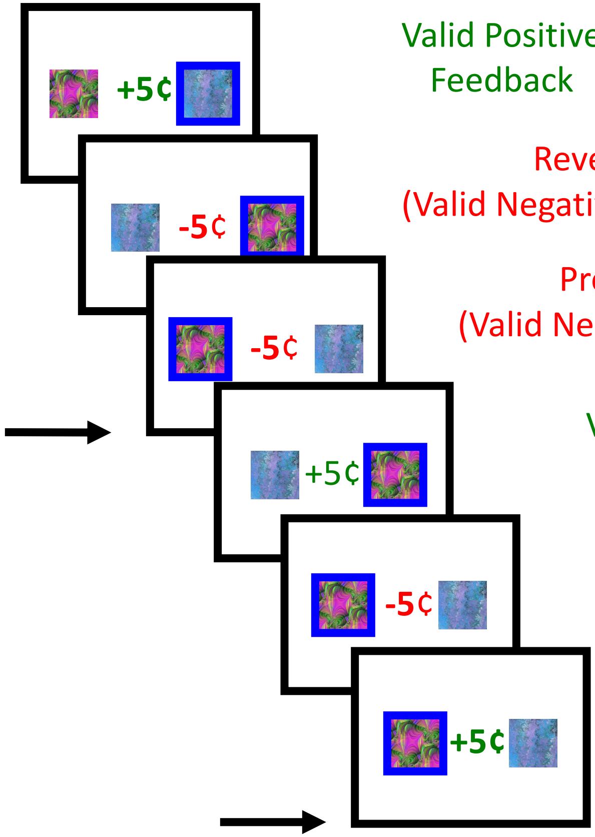
(Slightly more refined) Motivating Hypotheses

- Anhedonia and avolition in schizophrenia should be associated with abnormal reward prediction error signals

- Anhedonia and avolition in schizophrenia should be associated with abnormal reward anticipation signals, indicative of a reduced ability to assign incentive value to stimuli

Spoiler Alert: This is what we found.

Probabilistic Reversal Learning (PRL) Task



Valid Positive Feedback

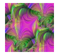
Reversal Error (Valid Negative Feedback/Shift)

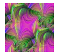
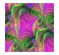
Pre-reversal Error (Valid Negative Feedback/Stay)

Valid Positive Feedback

Probabilistic Error (Invalid Negative Feedback)

Valid Positive Feedback

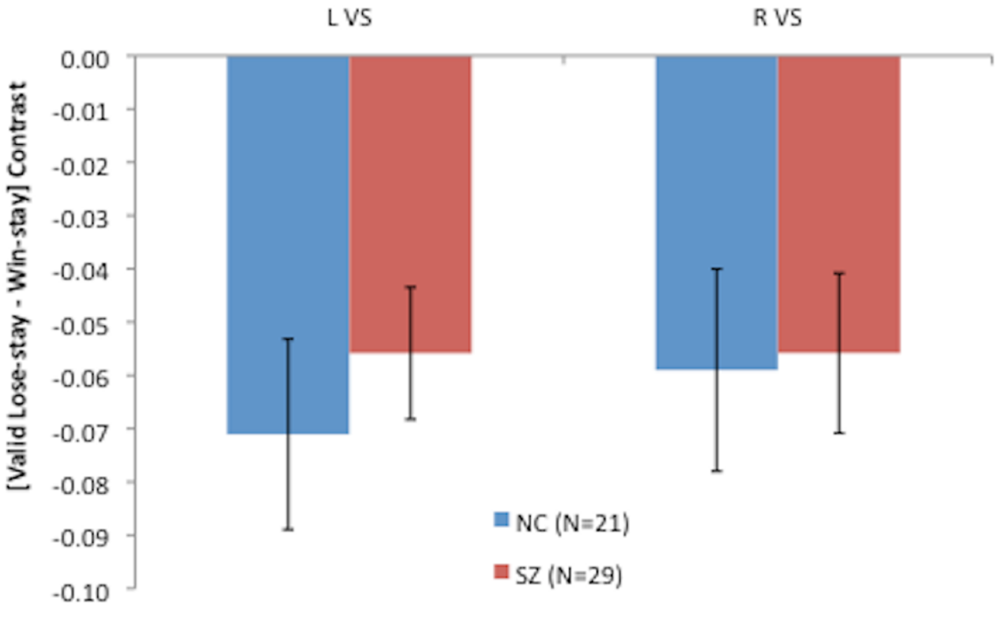
"Correct" stimulus switches
From  to

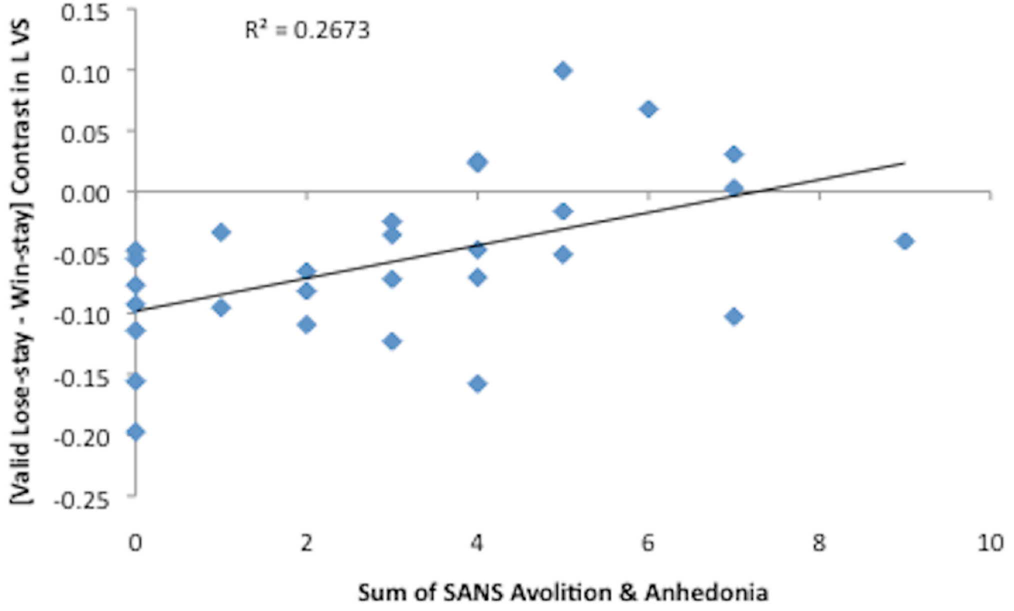
"Correct" stimulus is 
Choice of  is rewarded 8x out of 10.

PRL involves at least three processes:

1. Modulating attention, based on the salience of outcomes
2. Updating value representations based on violations of expectation (PEs)
3. Deciding based on expected values of choices

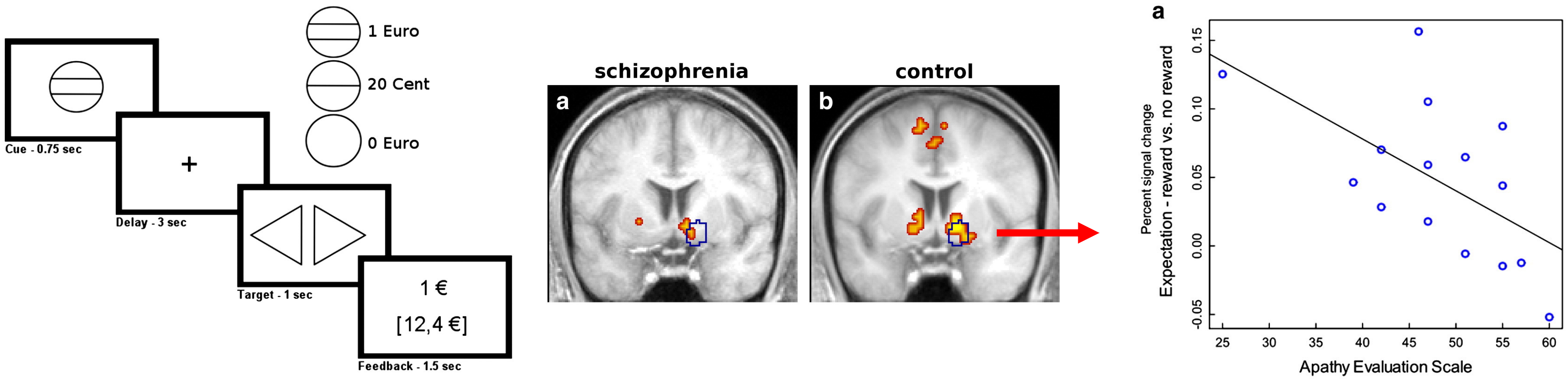
Striatal RPE signals have been shown to scale with ratings of anhedonia/avolition severity



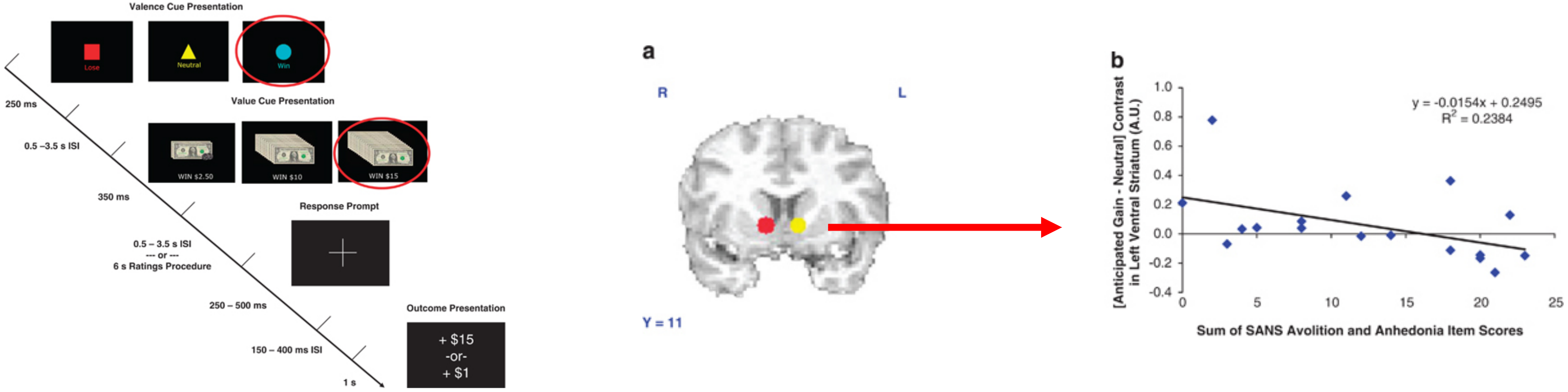


- The groups did not differ in their contrasts between expected and unexpected outcomes, but, in PSZ, contrasts between expected and unexpected outcomes correlated with ratings for negative symptoms

Striatal reward anticipation signals have been shown to scale with ratings of anhedonia/avolition severity



From Simon et al. (2010). Schiz. Res., 118, 154–161.



From Waltz et al. (2010). Neuropsychopharm., 35, 2427–2439.

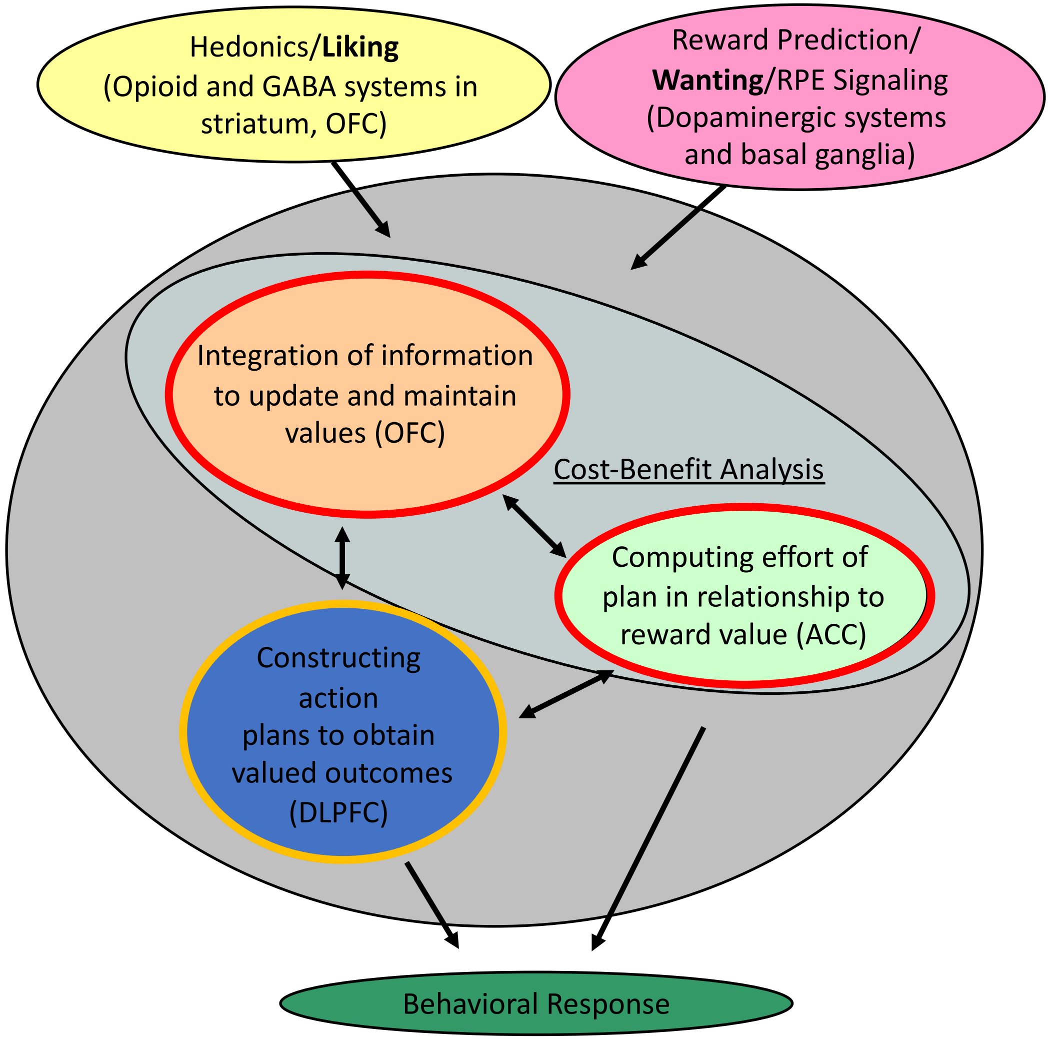
Q: Do Anhedonia and avolition in schizophrenia originate primarily with abnormal reward prediction error signals and abnormal reward anticipation signals (indicative of a reduced ability to assign incentive value to stimuli)>

A: They likely play a role, but there are many other ways in which goal-directed behavior can go wrong.

Goal is not just to find out what schizophrenia patients have difficulty with

- What is intact in most SZ patients?
- What is impaired only of SZ patients with motivational deficits?
- Multiple component processes involved in RL, including the signaling of the expected value (EV) of stimuli and actions, the integration of outcomes, and the signaling of reward prediction errors.
- Which reward-related signals could/do travel with clinical ratings of anhedonia/avolition?
- If avolition is not always driven by anhedonia (either consummatory or anticipatory), what is it driven *by*?
- What drives avolition in the presence of intact RPE signals?

There is more to goal-directed behavior than learning to want what you like


Hedonics/**Liking**
(Opioid and GABA systems in striatum, OFC)
Reward Prediction/ **Wanting**/RPE Signaling (Dopaminergic systems and basal ganglia)
Integration of information to update and maintain values (OFC)
Cost-Benefit Analysis
Computing effort of plan in relationship to reward value (ACC)
Constructing action plans to obtain valued outcomes (DLPFC)
Behavioral Response

From: Dowd and Barch (2011), After Wallis (2007)

Q: What did we think Computational Psychiatry could *buy* us?

A: A mechanistic account of avolition, through disrupted reinforcement learning and decision making.

What is necessary for learning about the value of stimuli and actions?

- Ability to integrate frequencies and magnitudes of potential outcomes

- Ability to represent both the costs and benefits of actions

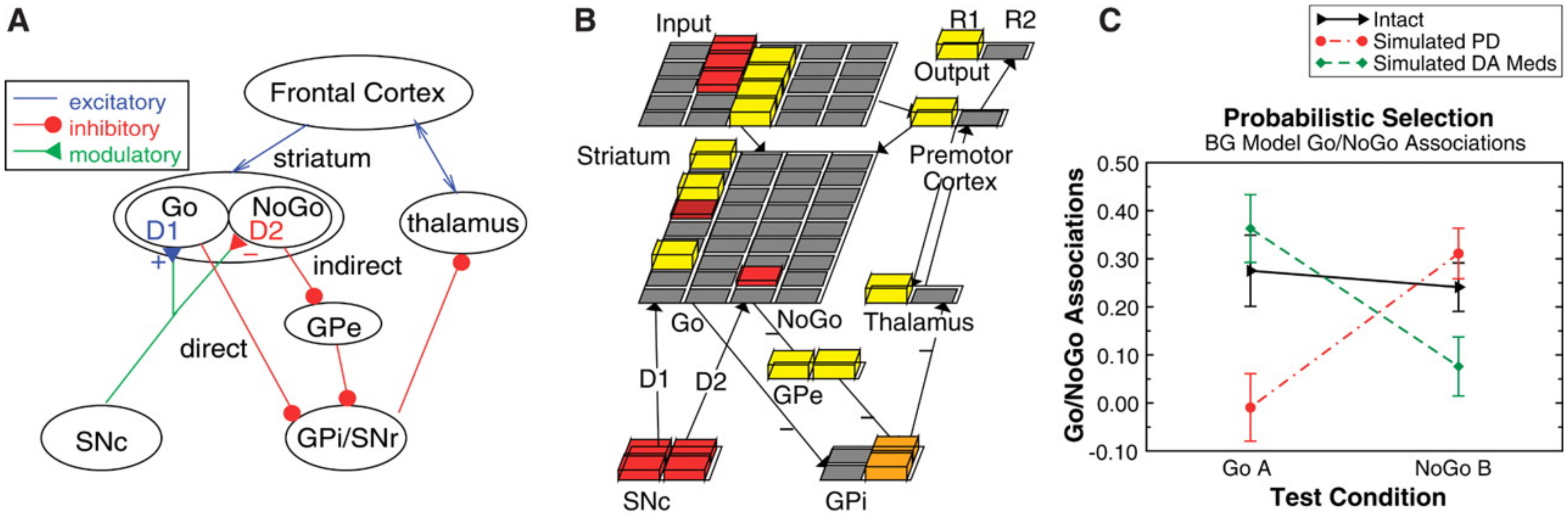
Kinds of RL

- Positive-RPE-driven- (Go-) vs. Negative-RPE-driven (NoGo-) Learning
- Rapid/PFC-driven/WM-dependent RL vs. Gradual/BG-driven/ Procedural RL
- Gain- vs. Loss-driven Learning
- Model-based vs. Model-free/State vs. Reward

Kinds of RL

- Positive-RPE-driven- (Go-) vs. Negative-RPE-driven (NoGo-) Learning
- Rapid/PFC-driven/WM-dependent RL vs. Gradual/BG-driven/Procedural RL
- Gain- vs. Loss-driven Learning
- Model-based vs. Model-free/ State vs. Reward

By Carrot or By Stick



The effect of a dopamine burst is to stimulate the D1/Direct/Go pathway and disinhibit the thalamus.

The effect of a dopamine dip is to release inhibition from the D2/Indirect/NoGo pathway and GPi, resulting in inhibition of the thalamus.

A. The corticostriato-thalamo-cortical loops, including the direct (Go) and indirect (NoGo) pathways of the basal ganglia. The Go cells disinhibit the thalamus via the internal segment of globus pallidus (GPi) and thereby facilitate the execution of an action represented in cortex. The NoGo cells have an opposing effect by increasing inhibition of the thalamus, which suppresses actions and thereby keeps them from being executed.

B. The Frank neural network model of this circuit (squares represent units, with height and color reflecting neural activity; yellow, most active; red, less active; gray, not active). The premotor cortex selects an output response via direct projections from the sensory input, and is modulated by the basal ganglia projections from thalamus.

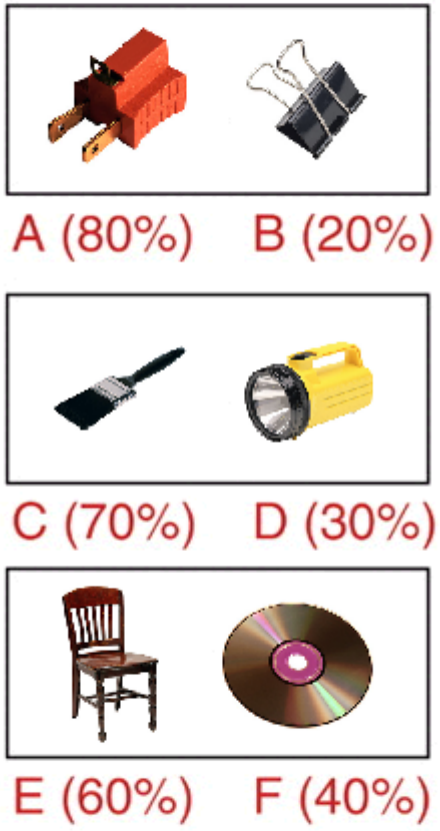
C. Predictions from the model for the probabilistic selection task, showing Go-NoGo associations for stimulus A and NoGo-Go associations for stimulus B. Error bars reflect standard error across 25 runs of the model with random initial weights.

Frank, M. J., Seeberger, L. C., & O’Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science,* 306(5703), 1940-1943.

Probabilistic Stimulus Selection Task

Acquisition/Training Phase

- 20 trials with each stimulus pair, per block
- 2-6 blocks
- Probabilistic feedback (“correct” or “incorrect”)
- Training ends when subject reaches criteria in all 3 conditions in same block
  - 65% A choices on AB trials
  - 60% C choices on CD trials
  - 50% E choices on EF trials
- Measures of Rapid/Declarative RL incude:
  - Performance on Training Pairs in first two blocks
  - Proportion of “wins” leading to “stays”
  - Proportion of “losses” leading to “shifts”

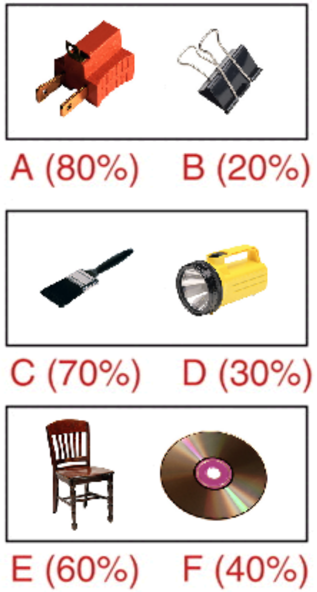


Probabilistic Stimulus Selection Task

Post-acquisition Test/Transfer Phase

- No feedback
- 4 trials with each of 15 possible stimulus pairing (60 total)

3 Training Pairs



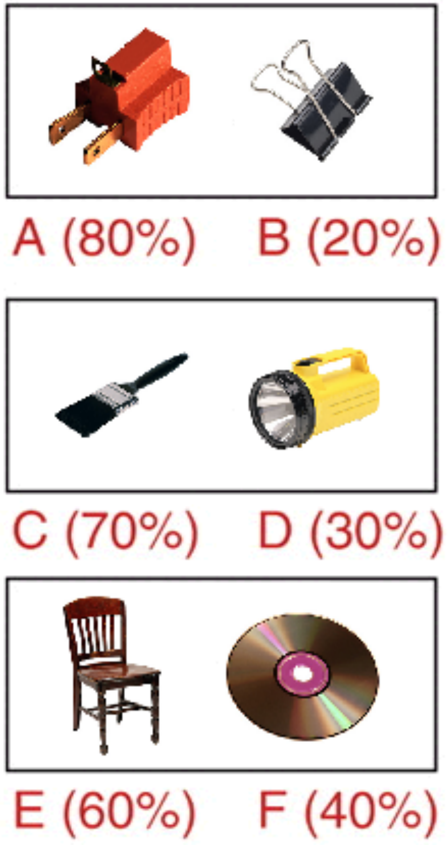
12 Novel Transfer Pairs

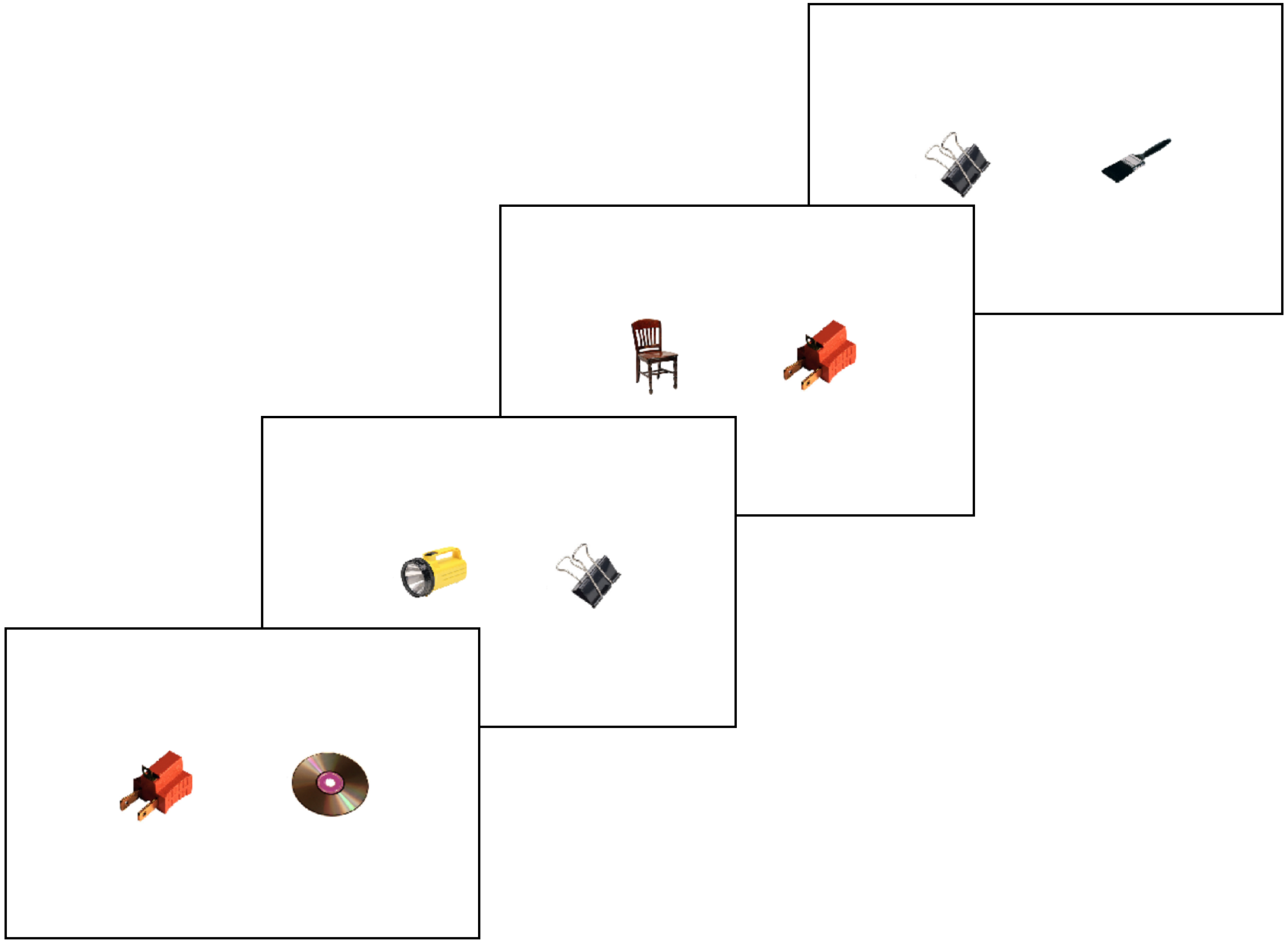
Choose A	Avoid B	Other
AC	BC	CE
AD	BD	CF
AE	BE	DE
AF	BF	DF

- “Go-learning” tested by transfer pairs with A (best)
- “NoGo-learning” tested by transfer pairs with B (worst)
- Have subjects learned the values of the best and worst stimuli?

Probabilistic Stimulus Selection Task: Post-acquisition Test Phase

Training Pairs



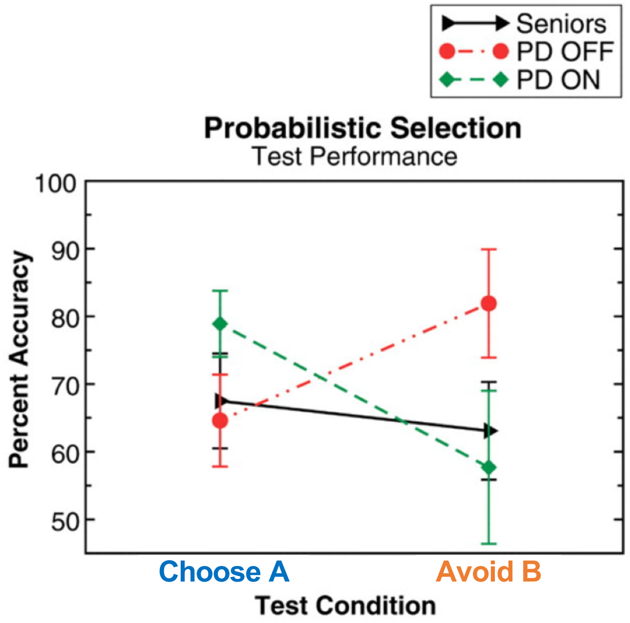


Go- and NoGo-learning in Parkinson’s Patients On and Off Dopamine Agonists

Transfer Pairs of Greatest Interest

Choose A	Avoid B
AC	BC
AD	BD
AE	BE
AF	BF

- “Go-learning” tested by transfer pairs with A (best)
- “NoGo-learning” tested by transfer pairs with B (worst)
- Have subjects learned the values of the best and worst stimuli?



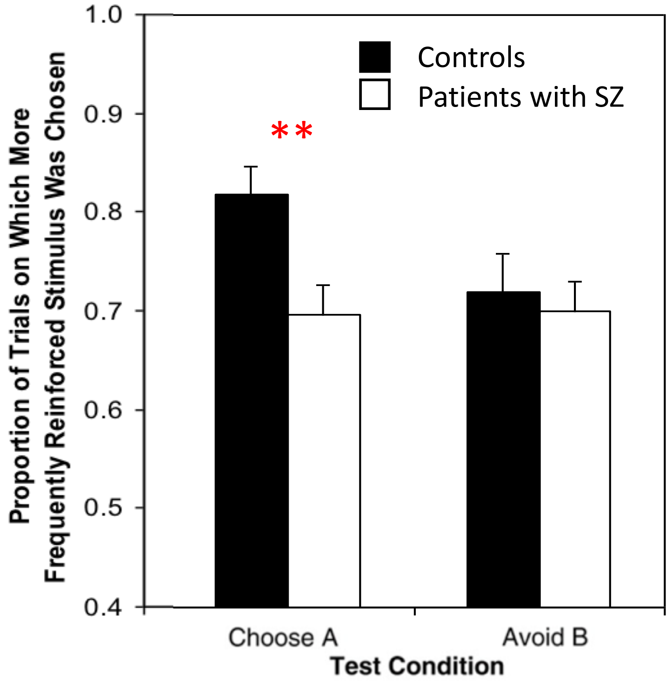
Dopamine depletion in Parkinson’s disease leads to reduce “Go-learning”, but enhanced “NoGo-learning.”

Frank, M. J., Seeberger, L. C., & O’Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science,* 306(5703), 1940-1943.

Go- and NoGo-learning in People with Schizophrenia

Predicted that faulty burst-firing in PSZ would lead to a Problem of Go-learning

- Patients would show impaired Choose-A behavior (Go-learning)
- Patients would show intact Avoid-B behavior (NoGo-learning)



** = p<0.01

Waltz et al. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry, 62,* 756-764.

But the story is more complicated than that…

There are many other RL mechanisms aside from BG-driven Go- and NoGo-learning

Kinds of RL

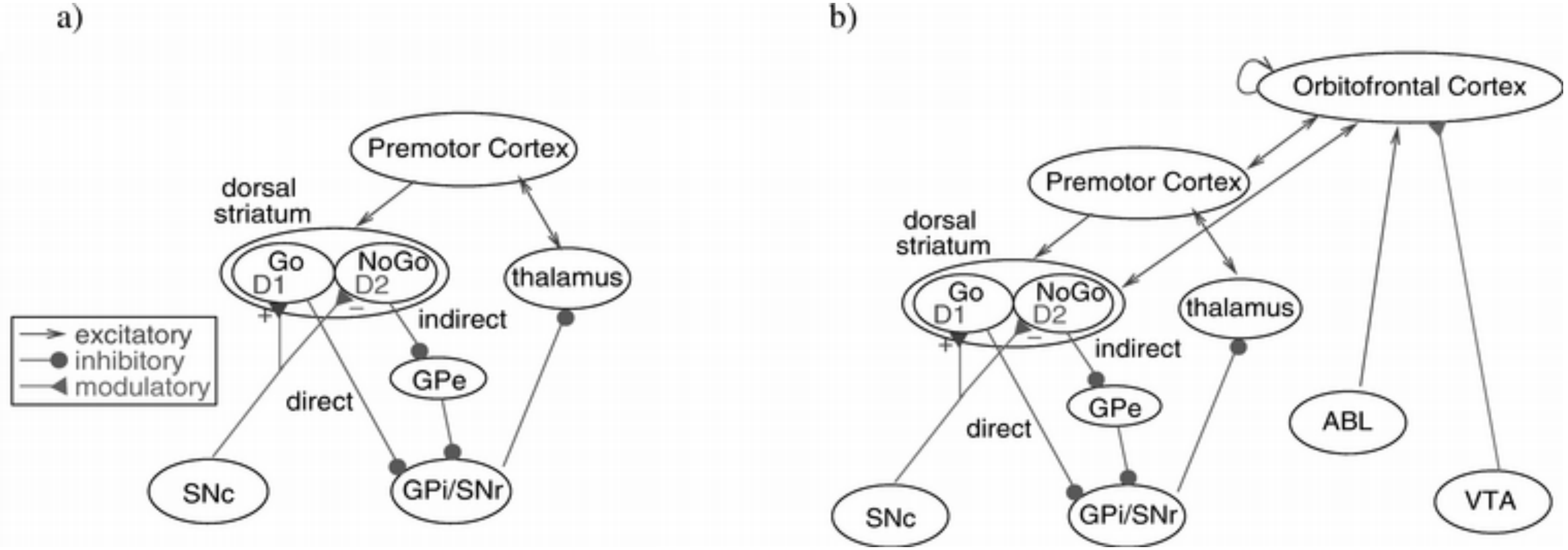
- Positive-RPE-driven- (Go-) vs. Negative-RPE-driven (NoGo-) Learning
- Rapid/PFC-driven/WM-dependent RL vs. Gradual/BG-driven/ Procedural RL
  - Evidence for *complementary* learning systems
- Gain- vs. Loss-driven Learning
- Model-based vs. Model-free

Rapid/PFC-driven/WM-dependent RL vs. Gradual/BG-driven/Procedural RL

- Dopaminergic RPE signals are thought to drive RL in the striatum, but semi-segregated D1 and D2 pathways not thought to drive Go- vs. NoGo-learning in the cortex

- Idea that orbitofrontal cortex (OFC) is there to represent the subjective value of stimuli, and that damage to OFC would lead to:
  - A reduced ability to precisely represent the magnitudes of outcomes (and flexibly modify these representations); and
  - A reduced ability to integrate the frequencies and magnitudes of outcomes
  - A reduced ability to learn over the course of 1 or 2 trials (win-stay and lose-shift)
  - A reduced ability to modify behavior in the face of sudden contingency reversals

Striato-Orbitofrontal Interactions and the integration of outcome frequency and magnitudes

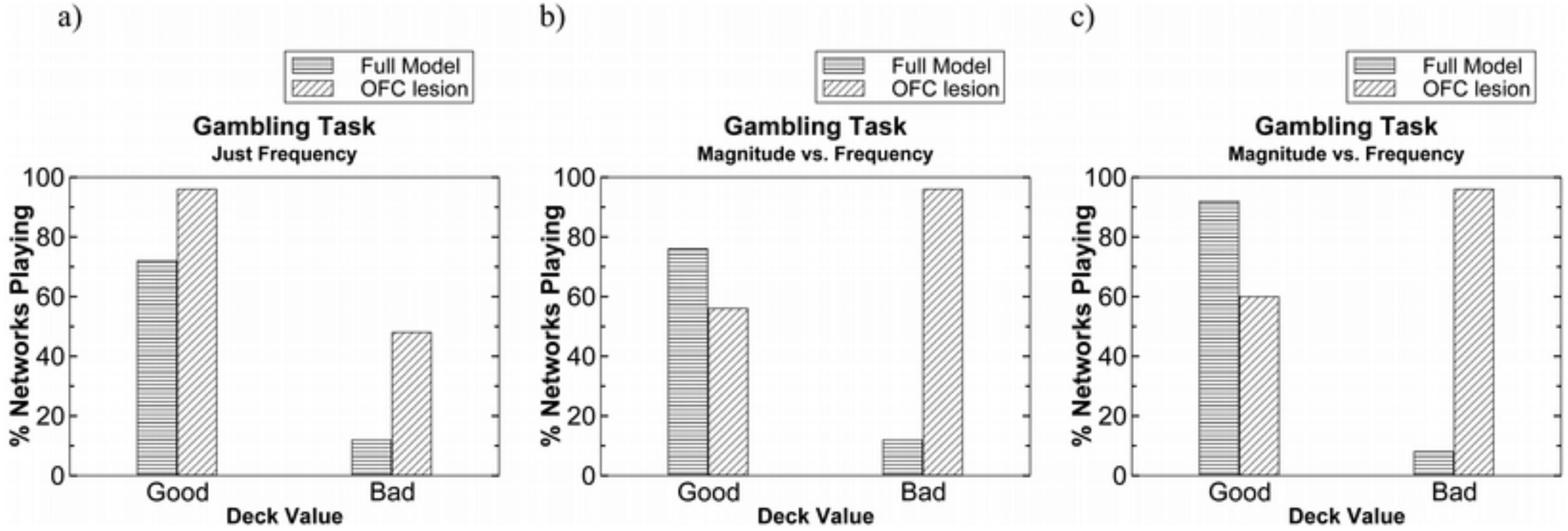


a) **The cortico-striato-thalamo-cortical loops, including the direct and indirect pathways of the basal ganglia.**

b) **The same circuit with additional influence from the orbitofrontal cortex, which can maintain reinforcement-related information in working memory and provide top-down biasing on the more primitive basal ganglia system, in addition to direct influencing of response selection processes in the premotor cortex**. The orbitofrontal cortex receives information about the relative magnitude of reinforcement values from the basolateral nucleus of the amygdala (ABL), which it can also maintain in working memory. Dopamine from the ventral tegmental area (VTA) projects to the ventral striatum (not shown) and the orbitofrontal cortex. GPe = external segment of the globus pallidus

Frank, MJ, Claus, ED. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review, 113,* 300-326.

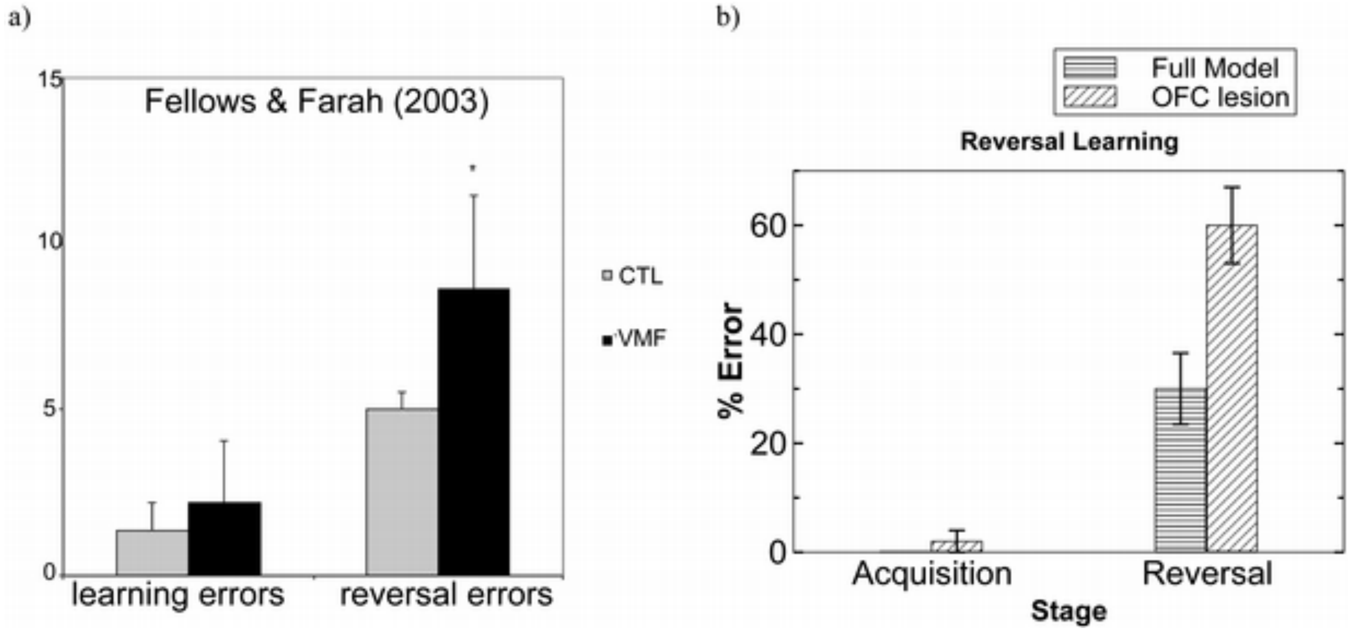
The integration of outcome frequencies and magnitudes enables one to solve the Iowa Gambling Task



- Gambling task results after 140 trials of training. a: In the just frequency condition, both intact and OFC-lesioned models were successful at playing to the good deck (which resulted in a gain 70% of the time) and passing on the bad deck (which resulted in a loss 70% of the time). b: When magnitude information was in opposition to frequency, the full model was nevertheless able to maximize expected value by playing on the infrequent high-gain deck and passing on the infrequent high-loss deck. In contrast, the OFC-lesioned networks continued to respond on the basis of frequency and therefore make maladaptive decisions. c: These results held up even when the dopamine signal was scaled such that high-magnitude gains-losses were associated with larger dopamine changes than were low-magnitude outcomes.

Frank, MJ, Claus, ED. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review, 113,* 300-326.

The integration of outcome frequencies and magnitudes enables one to detect rapid contingency reversals



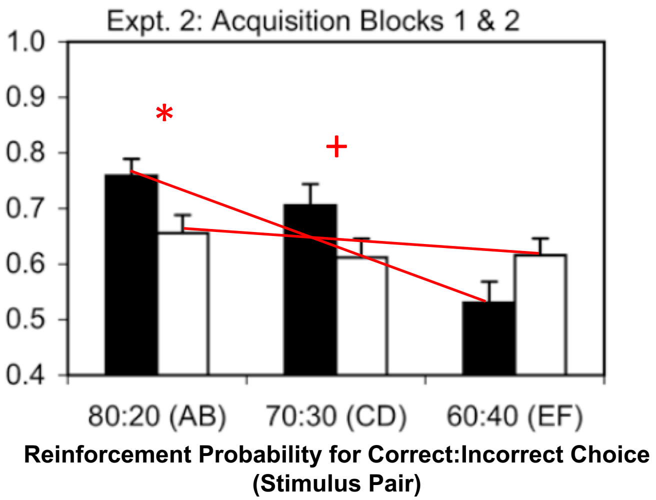
- a: Reversal learning impairments in humans with damage to ventromedial and orbitofrontal cortices, showing number of errors made in the learning and reversal phases. Modified from Fellows and Farah (2003) with permission. b: Model reversal learning results. Acquisition refers to performance (error percentages) after 200 trials; reversal refers to performance after a further 200 reversal trials.

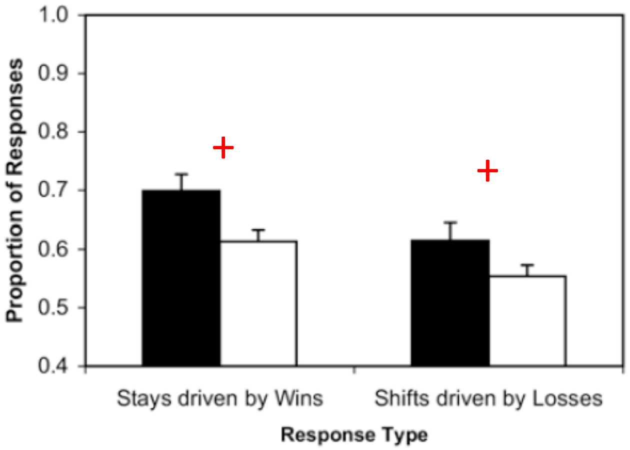
Frank, MJ, Claus, ED. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review, 113,* 300-326.

The integration of outcome frequencies and magnitudes enables one to modify behavior on a trial-wise basis and acquire reinforcement contingencies rapidly

Additional results from Probabilistic Stimulus Selection experiment:

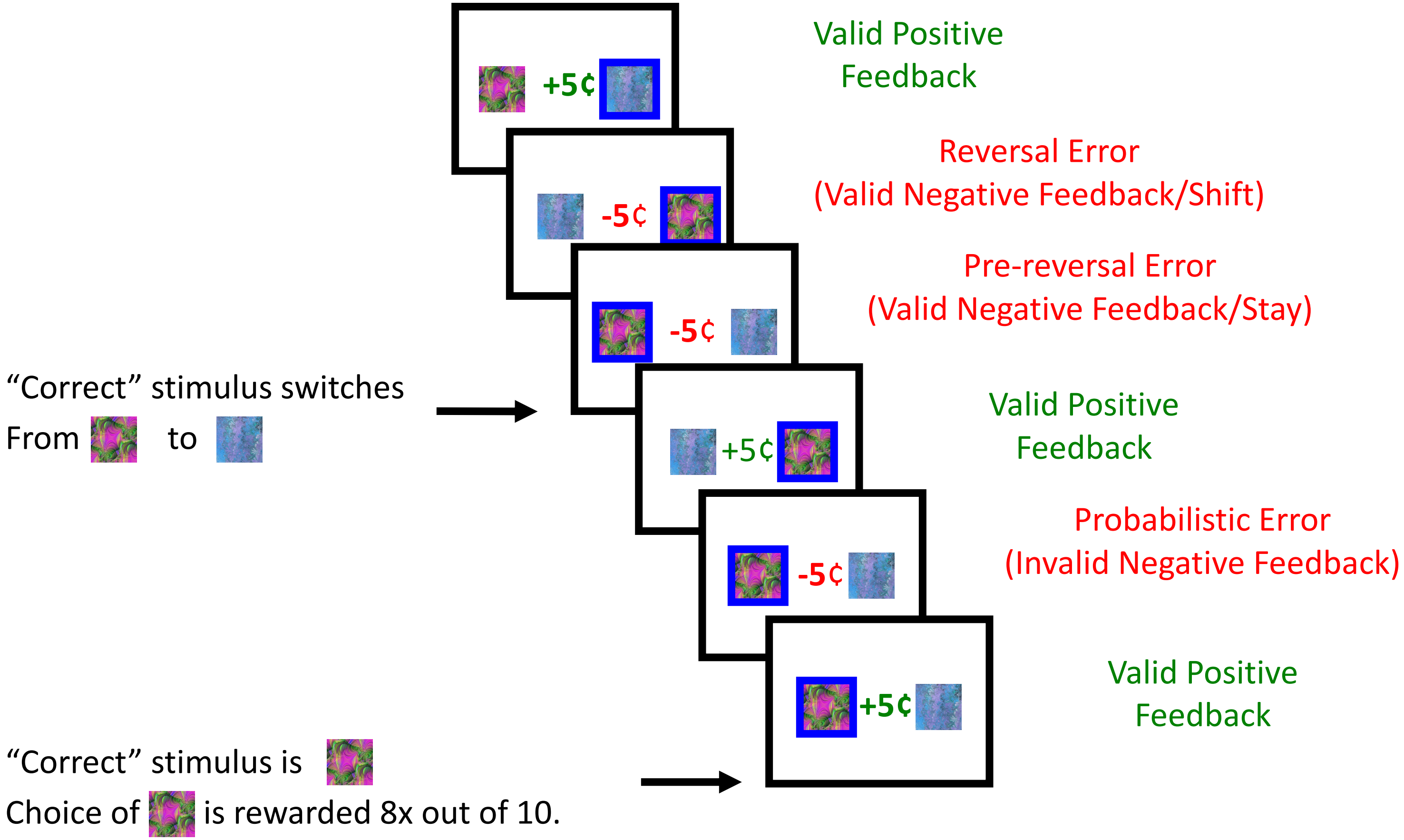
- During Acquisition, SZ patients showed reduced Early Acquisition scores on the higher-reward-frequency items (a measure of rapid RL), relative to controls
- Early Acquisition scores in SZ patients correlated with overall negative symptom scores
- During Acquisition, SZ patients showed reduced rates of both win-stay and lose-shift behavior – two other measures of rapid, trial-to-trial learning





Waltz et al. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry, 62,* 756-764.

Probabilistic Reversal Learning (PRL) as a Measure of the Ability to Detect Rapid Contingency Reversals



Valid Positive Feedback

Reversal Error (Valid Negative Feedback/Shift)

Pre-reversal Error (Valid Negative Feedback/Stay)

Valid Positive Feedback

Probabilistic Error (Invalid Negative Feedback)

Valid Positive Feedback

“Correct” stimulus switches
From 🟪 to 🟦

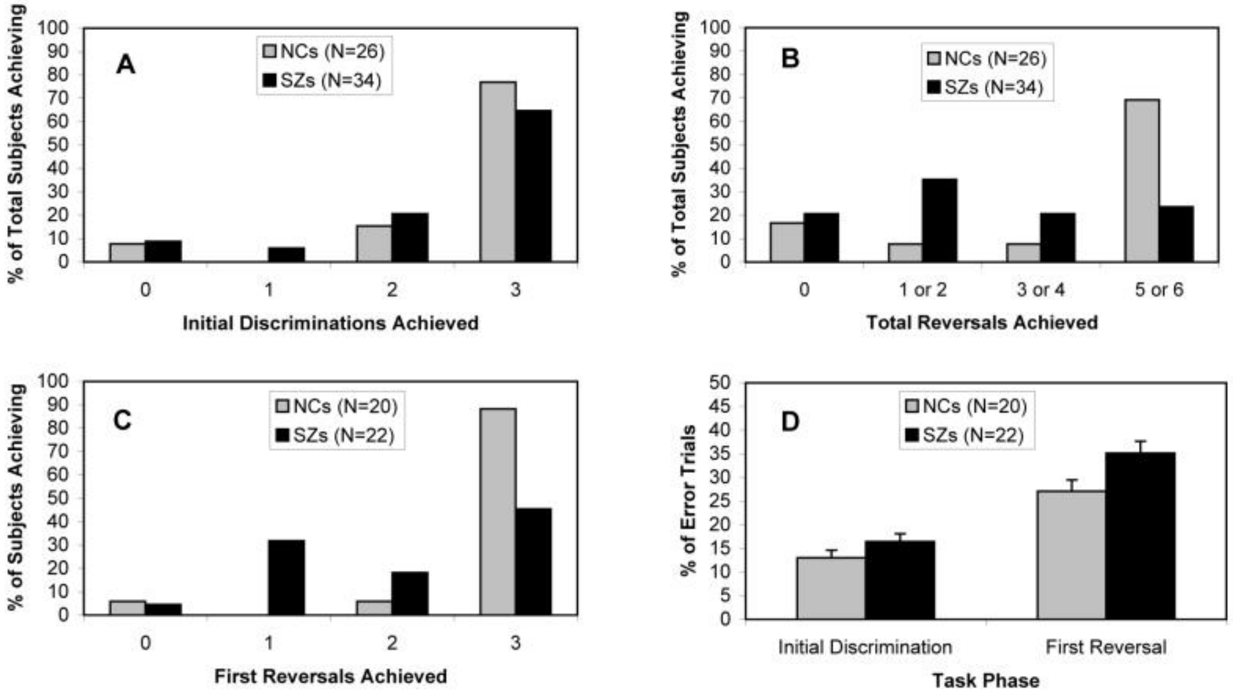
“Correct” stimulus is 🟪
Choice of 🟪 is rewarded 8x out of 10.

PRL involves at least three processes:

1. Modulating attention, based on the salience of outcomes
2. Updating value representations based on violations of expectation (PEs)
3. Deciding to repeat the previous response, or switch to the alternative response, based on expected values of choices, as well as certainty about the expected values of choices.

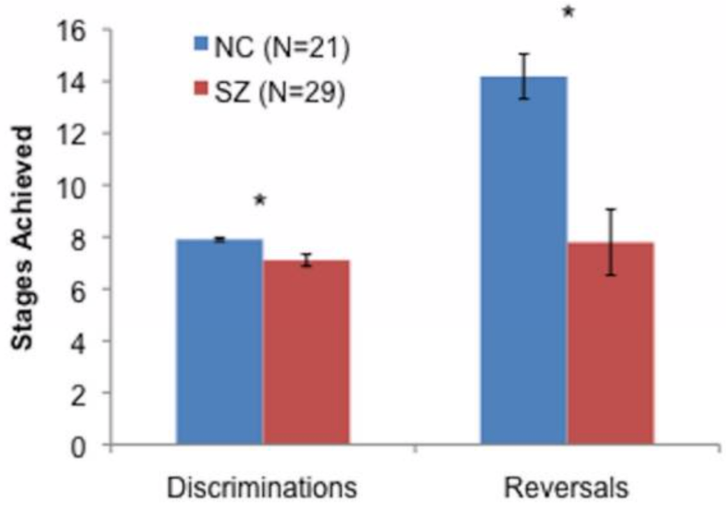
Probabilistic Reversal Learning as an Example of RL in an Unstable Environment

Experiment 1



Waltz et al. (2007). *Schizophrenia Research*.

Experiment 2

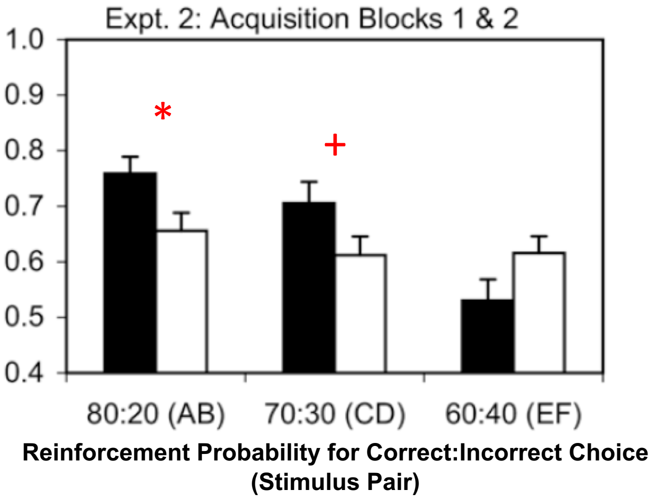


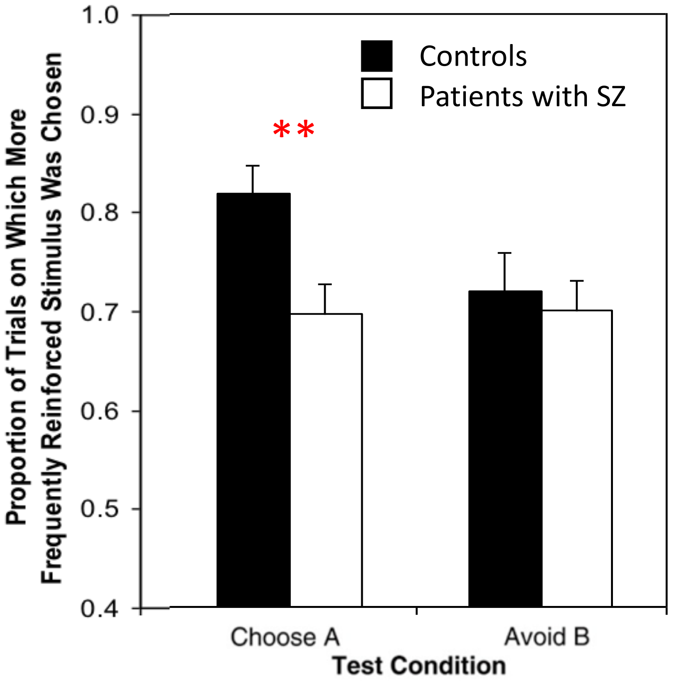
Waltz et al. (2013). *PLoS-ONE*.

- PSZ show much greater deficits in reversal of learned associations than in achievement of initial probabilistic discriminations

Gradation of RL Deficits in Schizophrenia

- Evidence that rapid/PFC-driven RL processes – esp. those involving +RPEs, are relatively more disrupted in SZ than slow/BG-driven RL processes (esp. those involving -RPEs), which may actually be somewhat preserved.





** = p<0.01

	Slow/ Procedural	Fast/ Declarative
Positive Feedback Driven	Basal Ganglia (D1)	VMPFC (D1)
Negative Feedback Driven	Basal Ganglia (D2)	VL/DMPFC (5HT)

Probabilistic Reinforcement Learning: Behavioral Phenomena We Have Linked to Avolition/Anhedonia in SZ

- Relatively-intact negative RPE-driven learning in the presence of impaired positive RPE-driven learning (Waltz et al. 2007; Waltz et al., 2011)

- Relatively-intact gradual/procedural learning in the presence of impaired rapid/explicit RL
  - Relatively-intact habit learning in the presence of impaired WM-dependent RL (Waltz et al. 2007; Waltz and Gold, 2007)
  - Greater performance deficits in SZ for more deterministic contingencies than less deterministic contingencies (value-difference effect; Hernaus et al., 2019a)

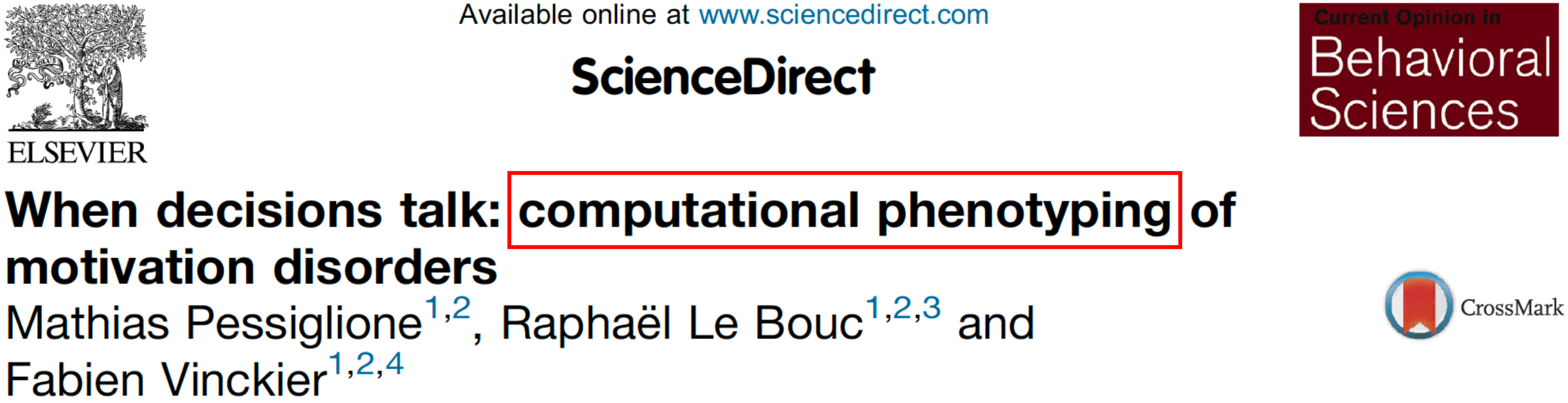
- Relatively-intact BG-driven learning in the presence of impaired OFC-driven RL
  - A reduced ability to integrate the frequencies and magnitudes of outcomes (Hernaus et al., 2019b)

- In general: a more limited ability to rapidly and flexibly update value representations in the brain (Waltz et al., 2015)

When we say that computational psychiatry can provide one with a mechanistic account of avolition, through disrupted RL and DM, we mean that it can generate mechanistic accounts of phenomena like these

And this is how we’ve tried to use computational methods…

Another potential benefit of computational psychiatry:



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Behavioral Sciences

When decisions talk: computational phenotyping of motivation disorders

Mathias Pessiglione[1,2], Raphaël Le Bouc[1,2,3] and Fabien Vinckier[1,2,4]

Types of Processes We’ve Examined

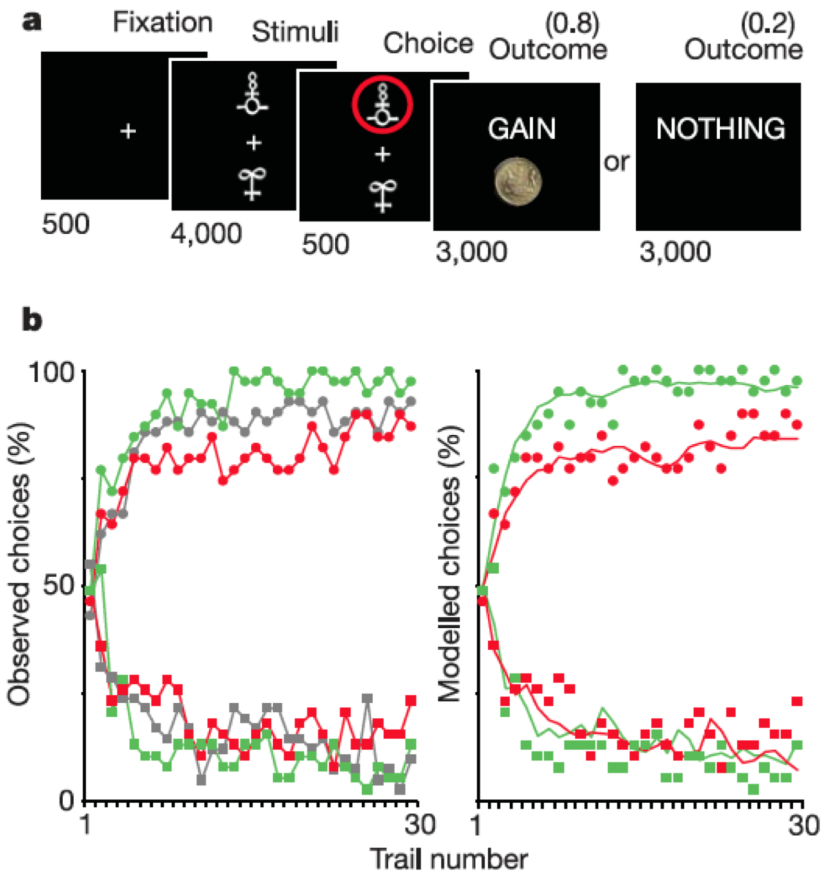
1. Acquisition vs. Expression of Learned Associations
2. Probabilistic RL in stable and unstable environments
3. Directed Exploration

II. How do we do what we do?

Considerations

- Are you modeling SR learning, or Q-learning, or both?
- How many learning rates, and which ones?
  - Separate learning rates for positive and negative RPEs? Actor and Critic?
- Decay parameter, or not?
  - Schonberg et al. (2007) sticky-choice model
- Explore parameter, or not?
- Who are your patients?
  - Are they old/young/medicated/unmedicated/at-risk/along the spectrum?
  - It should affect what you expect.
  - EVERYTHING I SAY TODAY, I BELIEVE TO BE TRUE OF MEDICATED ADULTS WITH (MULTIEPISODE) PSYCHOTIC ILLNESS, ESPECIALLY THOSE WITH MORE SEVERE NEGATIVE SYMPTOMS

Learning to seek gains and learning to avoid losses are likely at least semi-independent processes



- Haloperidol and L-dopa differentially affected RL, such that haloperidol affected reward-driven, but not punishment-driven RL

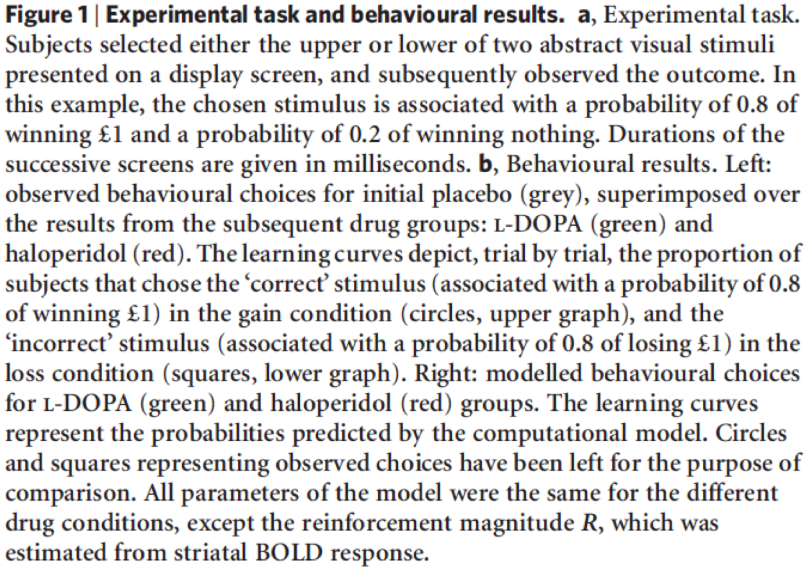


Figure 1 | Experimental task and behavioural results. a, Experimental task. Subjects selected either the upper or lower of two abstract visual stimuli presented on a display screen, and subsequently observed the outcome. In this example, the chosen stimulus is associated with a probability of 0.8 of winning £1 and a probability of 0.2 of winning nothing. Durations of the successive screens are given in milliseconds. **b**, Behavioural results. Left: observed behavioural choices for initial placebo (grey), superimposed over the results from the subsequent drug groups: L-DOPA (green) and haloperidol (red). The learning curves depict, trial by trial, the proportion of subjects that chose the 'correct' stimulus (associated with a probability of 0.8 of winning £1) in the gain condition (circles, upper graph), and the 'incorrect' stimulus (associated with a probability of 0.8 of losing £1) in the loss condition (squares, lower graph). Right: modelled behavioural choices for L-DOPA (green) and haloperidol (red) groups. The learning curves represent the probabilities predicted by the computational model. Circles and squares representing observed choices have been left for the purpose of comparison. All parameters of the model were the same for the different drug conditions, except the reinforcement magnitude *R*, which was estimated from striatal BOLD response.

Pessiglione, M., et al. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature,* 442 (7106), 1042-5.

Types of Models We Have Applied

- Q-learning models with one learning rate
- Q-learning models with more than one learning rate
  - Alpha-P vs. Alpha-N
- Actor/Critic models
- Hybrid models – to capture BG (slow) and OFC (fast) contributions
  - Q-learning + Actor/Critic
  - WM contribution
- Models with dynamic learning rates
- Models with exploration parameters

fMRI Data Analyses with Parametric Regressors

- Main goal is to identify increases and decreases in neural activity that are event-related
- In the context of tasks of reinforcement learning and decision making, our events of interest (cues and outcomes) can have variable amplitude:
    - Cues can have expected value and certainty about value
    - The value of outcomes can be expressed either in absolute terms, or as the difference between the expected and obtained outcomes
    - Trial-wise parameter estimates from RL models may serve as parametric/amplitude-modulated regressors
    - Event regressors are then convolved with an idealized hemodynamic response function
- NOW, when you perform group-level analyses (t-tests, ANOVAs, mixed-effect analyses), the beta coefficients from single-subjects regression analyses reflect the extent to which fluctuations in brain activity track internal representations of the values we are modeling (like expected value, certainty about value, and reward prediction errors)

Combining behavioral modeling with neuroimaging

1. Pass individual subject trial history to model
2. Find best-fitting parameters of model to behavioral data
3. Generate model-based time series
4. Convolve time series with hemodynamic response function
5. Regress against fMRI data

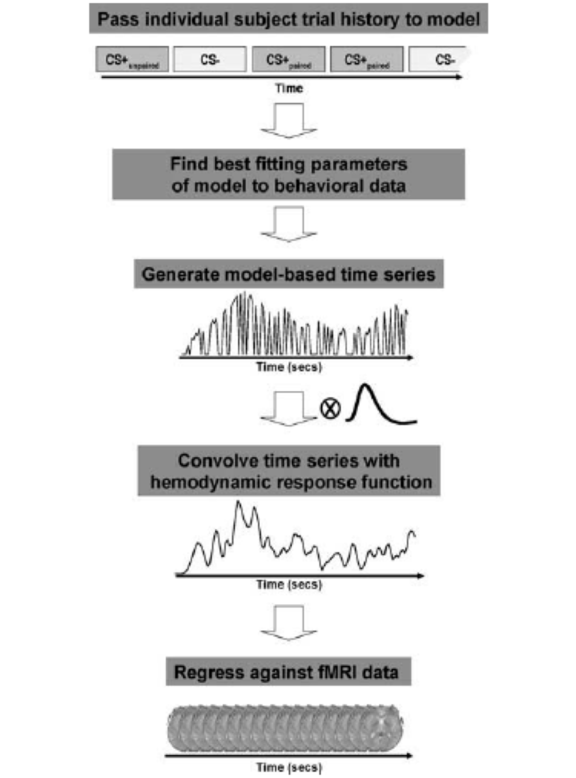

Pass individual subject trial history to model
CS+unpaired
CS-
CS+paired
CS+paired
CS-
Time
Find best fitting parameters of model to behavioral data
Generate model-based time series
Time (secs)
⊗
Convolve time series with hemodynamic response function
Time (secs)
Regress against fMRI data
Time (secs)

FIGURE 1. Illustration of model-based fMRI approach. Each individual subject's trial history is passed to the model, and the parameters of the model are fit so as to minimize the difference between the model predictions and an external behavioral measure, which in the conditioning example could be an external measure of conditioning, such as galvanic skin conductance responses or pupil dilation. Next, the best model-fitting parameters are used to generate a time series for each trial in the fMRI, which are then convolved with basis function(s) to account for the effects of hemodynamic lag, such as the canonical hemodynamic response function, and then regressed against the fMRI data.

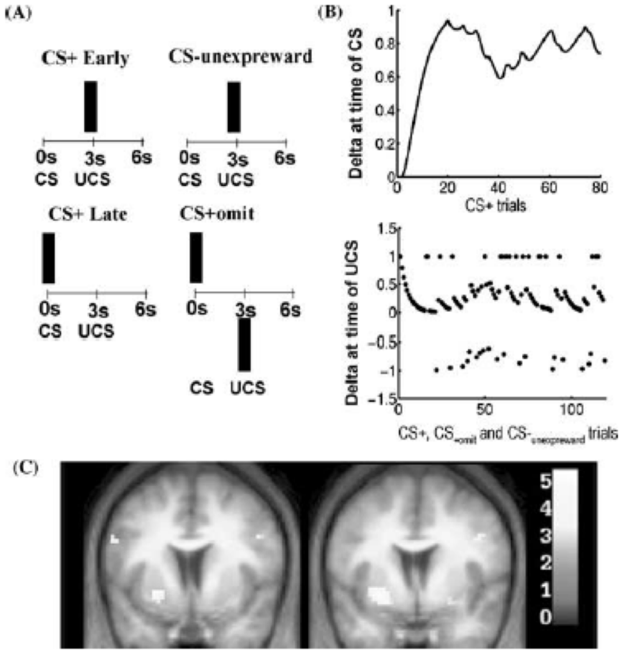


FIGURE 2. Model-based fMRI of stimulus-reward learning. (**A**) Properties of the temporal difference prediction error signal during reward learning in which a cue (CS+) is paired repeatedly with a reward (UCS) presented 3 sec later. During the initial stages of learning (CS + early trials), the error signal responds at the time of presentation of the UCS, but over the course of learning transfers back to the time of presentation of the CS (CS + late trials). On trials in which the CS+ is not presented but the reward is delivered anyway (CS–unexp. reward), the signal shows a positive response at the time the reward is delivered, whereas on trials in which the CS is presented but the reward is unexpectedly omitted the signals show a negative response at the time of outcome. (**B**) Plot of model-generated prediction error signals at the time of presentation of the CS, and the time of presentation of the UCS, over the course of the experiment for a typical subject. (**C**) Area of bilateral ventral striatum (ventral putamen bilaterally) showing significant correlations with the temporal difference prediction error signal while subjects underwent classical conditioning with sweet taste reward (1M glucose). Data from O'Doherty *et al.*[11]

O'Doherty, J.P., et al. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. *ANYAS, 1104,* 35–53.

III. Modeling Probabilistic RL in a Stable Environment

Probabilistic Reinforcement Learning:
Behavioral Phenomena We Have Linked to Avolition/Anhedonia in SZ

- Relatively-intact negative RPE-driven learning in the presence of impaired positive RPE-driven learning

- **Relatively-intact gradual/BG-driven/procedural learning in the presence of impaired rapid/OFC-driven/explicit RL**
  - Relatively-intact habit learning in the presence of impaired WM-dependent RL
  - Greater performance deficits in SZ for more deterministic contingencies than less deterministic contingencies (value-difference effect)
  - A reduced ability to integrate the frequencies and magnitudes of outcomes

- In general: a more limited ability to rapidly and flexibly update value representations in the brain

Why would a system with intact signaling of RPEs fail to adaptively represent the values of choices?

If it didn’t have a Q-learning mechanism, allowing it to represent the magnitudes of outcomes and integrate them with representations of the frequencies of outcomes

Modified Pessiglione Probabilistic Selection Task/ Gain vs. Loss-avoidance (GLA) Task

- Allows one to cross the valence of the outcome (gain/loss/neutral) with the valence of the prediction error (positive or negative), allowing one to perform contrasts across levels of outcome valence (gain vs. neutral, e.g.), for the same level of RPE (when both are better than expected), as well as perform contrasts across levels of RPE valence (positive vs. negative), for the same level of outcome (neutral, e.g.)

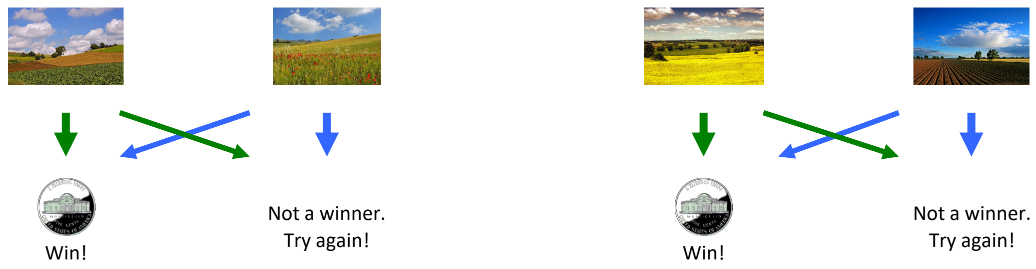
- Allows one to represent both the acquisition of contingencies and the expression of learned contingencies
- One can also use a computational model to estimate expected value and RPE on a trial-wise basis, to determine how well different brain regions track RPE valence and magnitude through their activity

- One can ask:
  - Do (avolitional) SZ patients show aberrant neural signals for all forms of positive and negative RPEs?
  - Do (avolitional) SZ patients show a specific abnormality in signaling the occurrence of gains, relative to losses, or even relative to instances of loss-avoidance – another kind of positive prediction error

Gain vs. Loss-avoidance (GLA) Task: Acquisition Phase

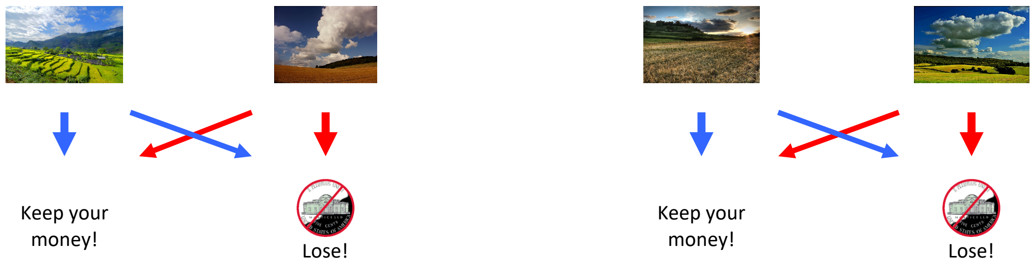
160 trials (40 with each of 4 pairs), with monetary feedback

2 Gain/Miss pairs (AB, CD): Frequent Winners (FW) vs. Infrequent Winners (IW)



	Frequent (90%) Winner (A)	vs.	Infrequent (10%) Winner (B)	Frequent (80%) Winner (C)	vs.	Infrequent (20%) Winner (D)
Stimuli						
Possible Outcomes	Win!		Not a winner. Try again!	Win!		Not a winner. Try again!
Expected Value	.9 x .05 = 45 cents		.1 x .05 = 5 cents	.8 x .05 = 40 cents		.2 x .05 = 10 cents

2 Loss/Avoid pairs (EF, GH): Infrequent Losers (IL) vs. Frequent Losers (FL)



	Infrequent (10%) Loser (E)	vs.	Frequent (90%) Loser (F)	Infrequent (20%) Loser (G)	vs.	Frequent (80%) Loser (H)
Stimuli						
Possible Outcomes	Keep your money!		Lose!	Keep your money!		Lose!
Expected Value	.1 x -.05 = -5 cents		.9 x -.05 = -45 cents	.2 x -.05 = -10 cents		.8 x -.05 = -40 cents

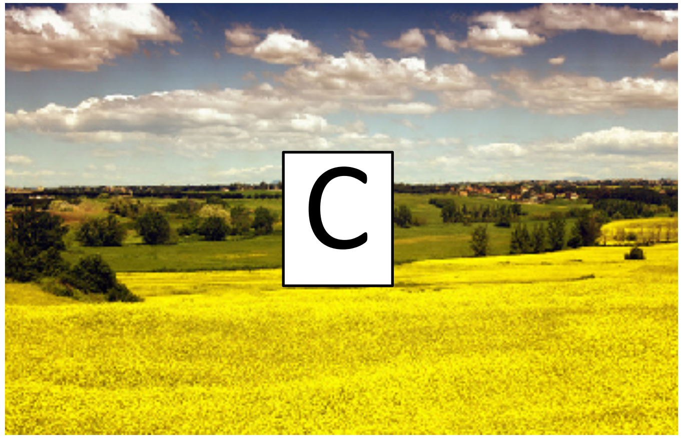
Gold et al. (2012). Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Arch. Gen. Psychiat., 69,* 129-38.

B

E PLURIBUS UNUM
MONTICELLO
FIVE CENTS
UNITED STATES OF AMERICA

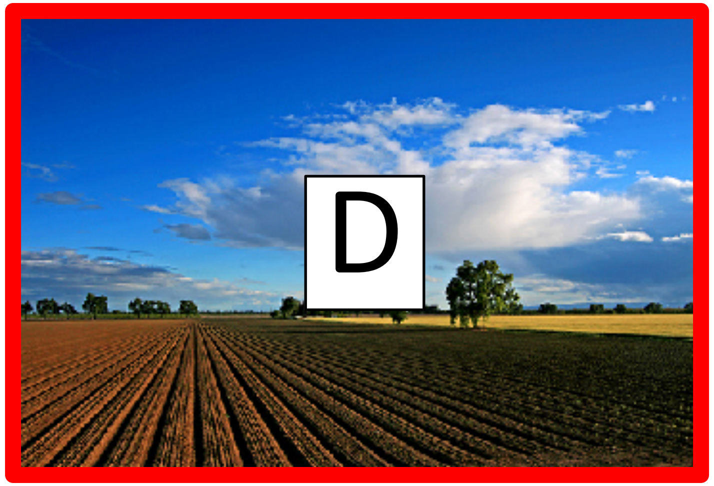
Win!

A



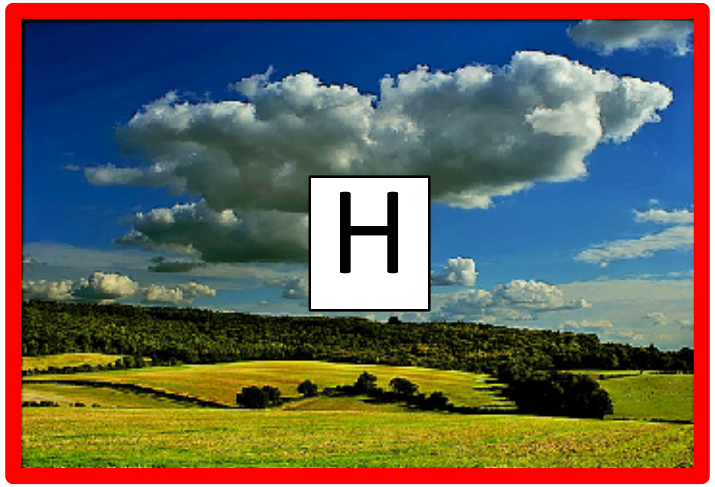
C

Not a winner.
Try again!



D

Keep your money!

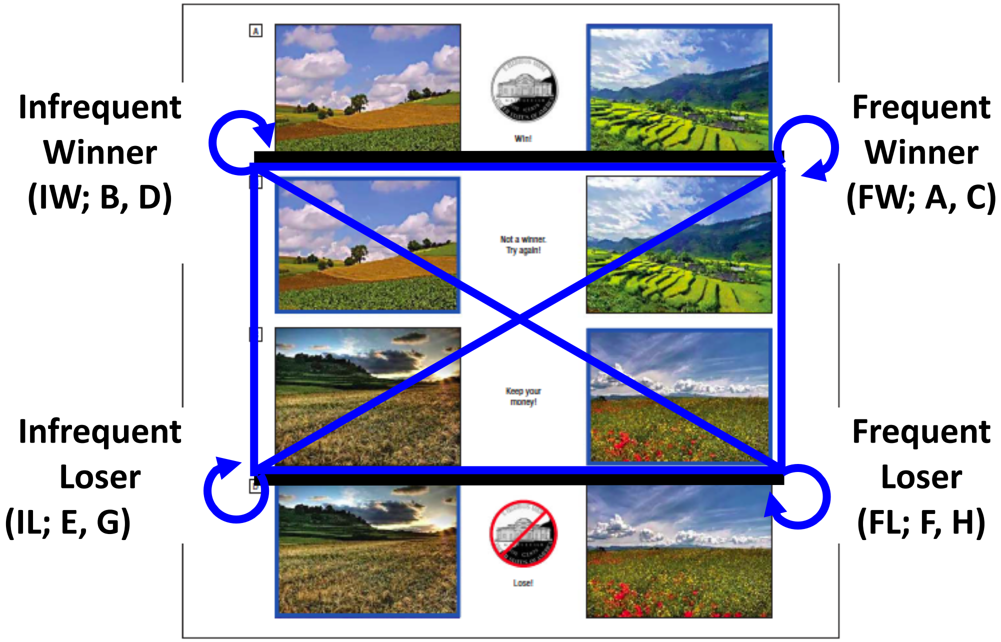


Lose!

Prob. Selection/Gain vs. Loss-avoidance Task (GLAT)

Transfer Phase: 64 trials, with all possible stimulus combinations, and no feedback

- The 4 training pairs were each presented 4 times (16 total trials);
- 48 trials with 10 kinds of novel pairings



Infrequent Winner (IW; B, D)

Frequent Winner (FW; A, C)

Infrequent Loser (IL; E, G)

Frequent Loser (FL; F, H)

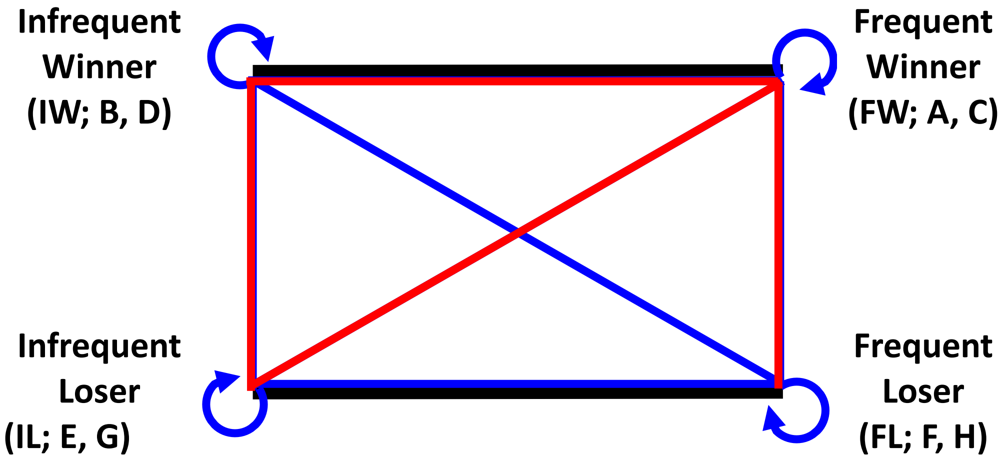
Figure 1. Example of reinforcement learning task stimuli and feedback. A, Feedback delivered after a correct choice (indicated by a blue border) in the reward trials. B, Feedback delivered following an incorrect choice. C, Feedback delivered following a correct choice in the loss-avoidance trials. D,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

Gold et al. (2012) Arch. Gen. Psychiat.

Prob. Selection/Gain vs. Loss-avoidance Task (GLAT)

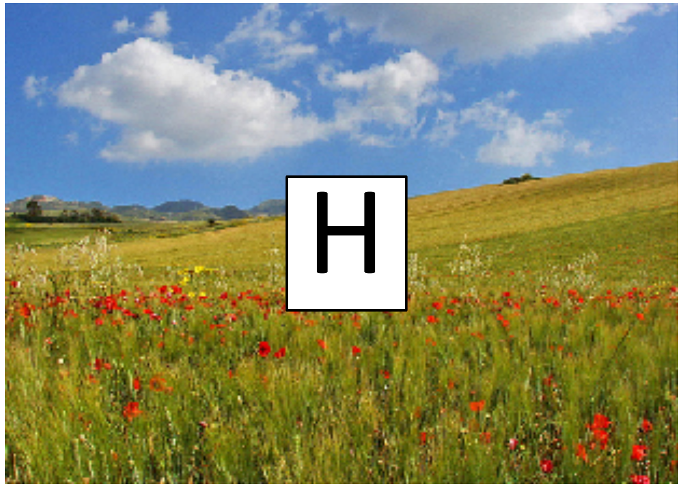
Transfer Phase: 64 trials, with all possible stimulus combinations, and no feedback

- The 4 training pairs were each presented 4 times (16 total trials);
- 48 trials with 10 kinds of novel pairings

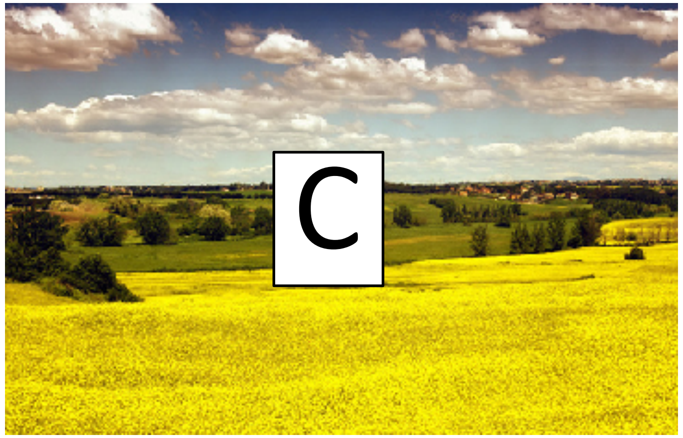


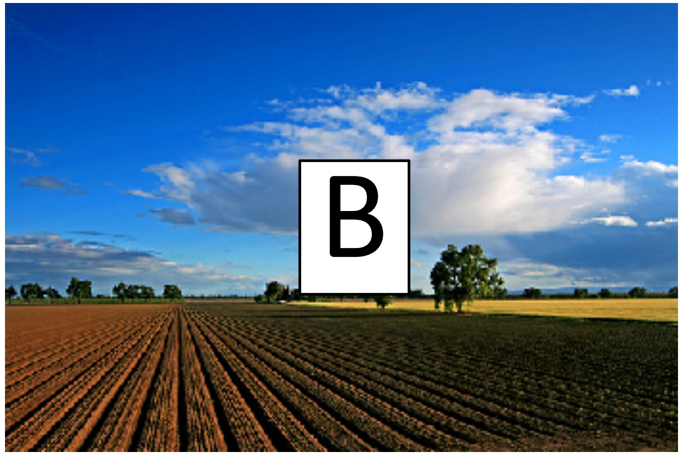
- Transfer contrasts of interest
  - FW vs. FL (+PE Gain vs. -PE Loss)
  - FW vs. IW (+PE Gain vs. -PE Neutral)
  - IL vs. IW (+PE Neutral vs. -PE Neutral)
  - FW vs. IL (+PE Gain vs. +PE Neutral)

Gold et al. (2012) Arch. Gen. Psychiat.


H

A


C


B

E

D

A

E

Gain vs. Loss-avoidance Task: Acquisition

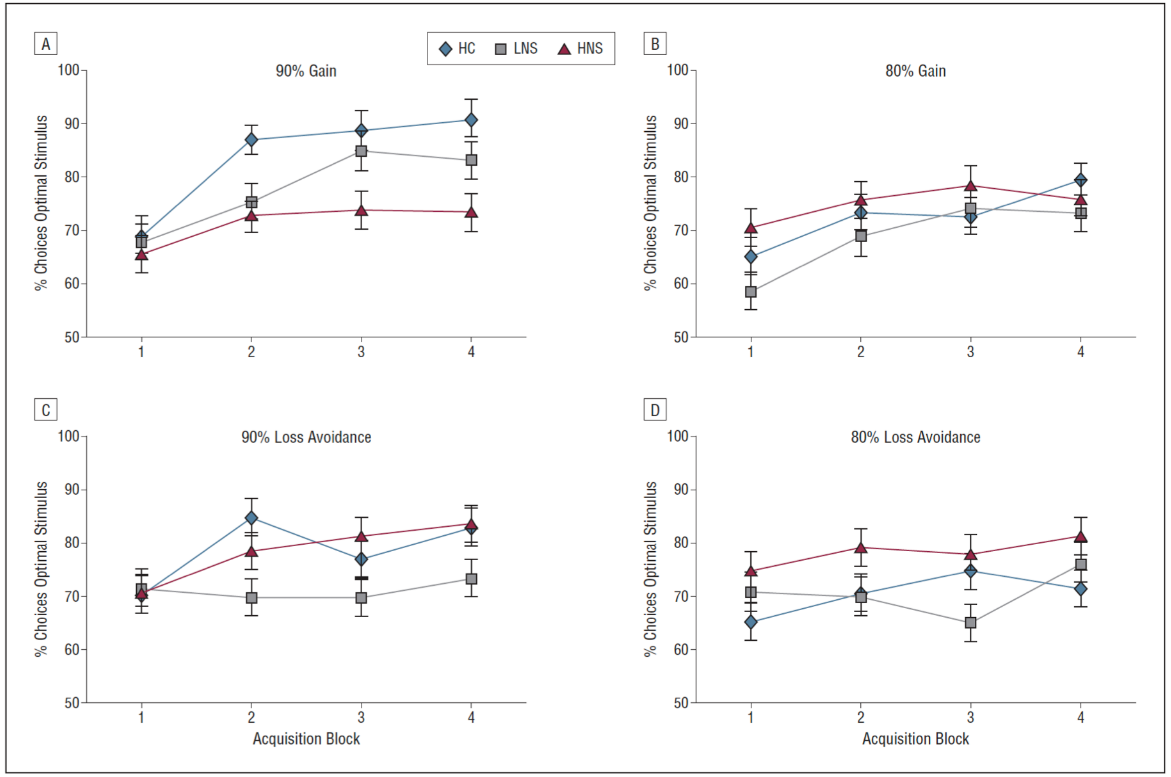


Figure 2. Differences in reinforcement learning among patients and healthy control (HC) subjects in 90% and 80% probability gain and loss-avoidance conditions. A and B, Performance in the 90% and 80% gain conditions, respectively. C and D, Performance in the 90% and 80% loss-avoidance conditions, respectively.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

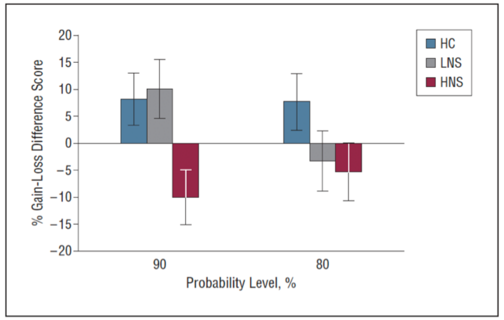
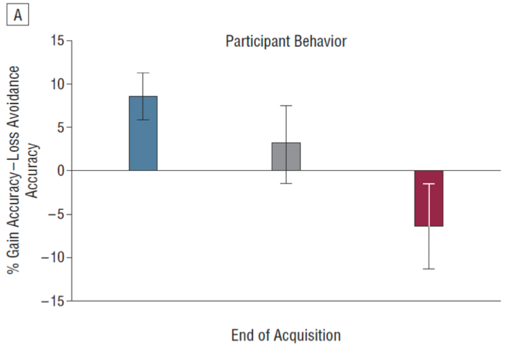


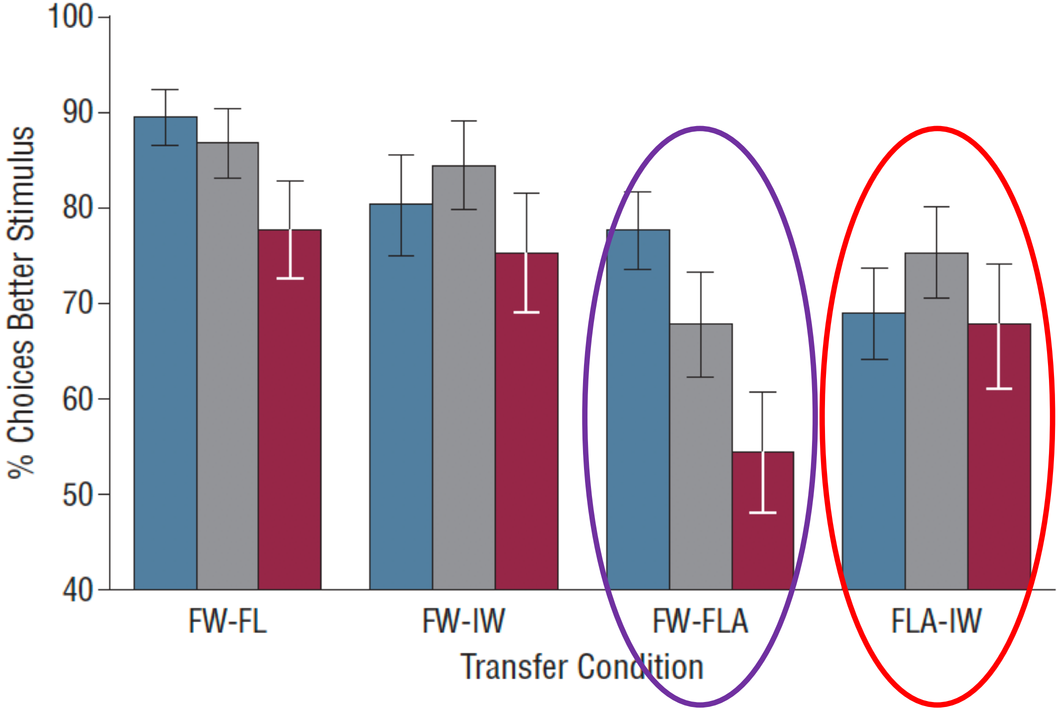
Figure 3. Performance on the gain and loss-avoidance difference score among patients and healthy control (HC) subjects. The difference score was calculated using block 4 performance. Scores above zero indicate better learning from gain than from loss avoidance, while scores below zero indicate better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



- Controls and patients with less severe negative symptoms show an advantage for gain-driven learning over loss-avoidance driven learning
- Patients with more severe negative symptoms show the opposite; they are better at learning what *not* to do than what *to* do…

Gold et al. (2012). Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Arch. Gen. Psychiat., 69,* 129-38.

Gain vs. Loss-avoidance Task: Transfer Results



Two effects of particular interest:

- All groups preferred Frequent Loss-avoiders to Infrequent Winners
- PSZ with more severe anhedonia/avolition like Frequent Loss-avoiders as much as they like Frequent Winners

Gold et al. (2012). Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Arch. Gen. Psychiat., 69,* 129-38.

Actor-Critic vs. Q-learning

- Unlike a Q-learning model, an **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.
  - Critic's Expected Value and the action weight are represented and updated separately

- Good at capturing S-R learning phenomena, habit learning

- Participants update the expected value V(t) of a state context on each trial t, according to the prediction error

- ε(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.

$$V(s,t+1) = V(s,t) + \alpha_C * \varepsilon(t),$$

- Prediction errors in the critic are, ε(t), is also used to update the stimulus-response weight, w(s,a,t), for the action selected in trial t

$$w(s,a,t+1) = w(s,a,t) + \alpha_A * \varepsilon(t),$$

$$w(s,a_1,t) \leftarrow w(s,a_1,t) / (|w(s,a_1,t)| + |w(s,a_2,t)|).$$

- Actions are selected according to the standard softmax logistic function:

$$P(a_1,t)= e^{(w(s,a_1,t)/\beta)} / (e^{(w(s,a_1,t)/\beta)} + e^{(w(s,a_2,t)/\beta)}),$$

Actor-Critic vs. Q-learning (cont.)

- Q-learning model does learn specific state-action values:

$$Q(a,t+1) = Q(a,t) + \alpha_O{}^*(outcome(t) - Q(a,t)),$$

- Actions again selected according to the standard softmax logistic function:

$$P(a_1,t)= e^{(Q(a_1,t)/\beta)} \,/\, (e^{(Q(a_1\,,t)/\beta)} + e^{(Q(a_2,t)\,/\beta)}),$$

- This, time, actions are chosen according to values, not weights

The Hybrid Model combines Actor-Critic and Q-learning mechanisms using a mixing parameter

$$(1)\ Q_t(\mathrm{s},\mathrm{a}) = Q_{t-1}(\mathrm{s},\mathrm{a}) + \alpha_Q * \delta(\mathrm{t})$$

$$(2)\ V_t(\mathrm{s}) = V_{t-1}(\mathrm{s}) + \alpha_C * \delta(\mathrm{t})$$

$$(3)\ W_t(\mathrm{s},\mathrm{a}) = W_{t-1}(\mathrm{s},\mathrm{a}) + \alpha_A * \delta(\mathrm{t})$$

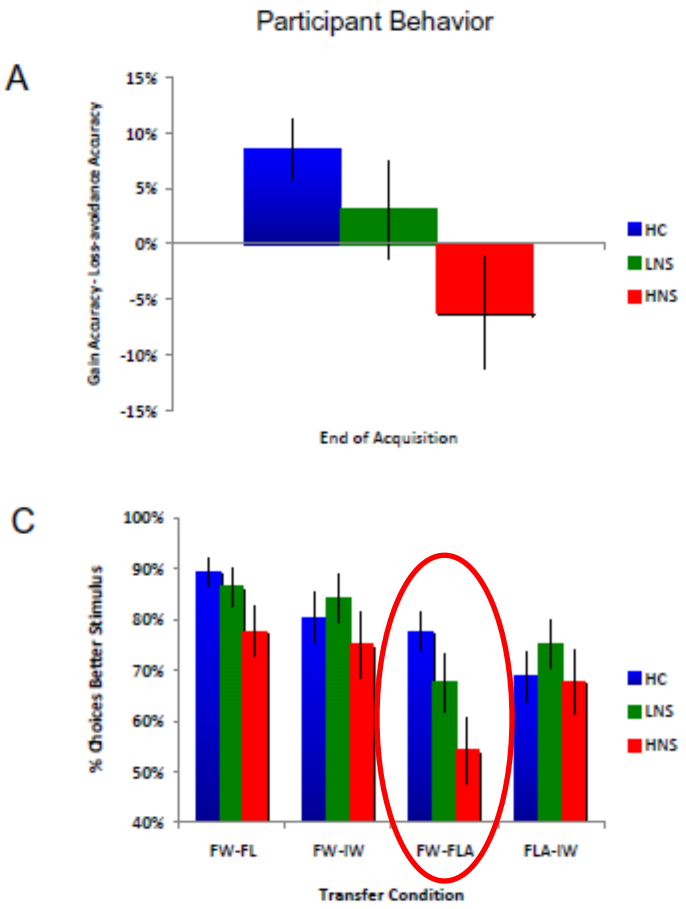
$$(4)\ Q\_AC_t(\mathrm{s},\mathrm{a}) = \left((1 - \boxed{m}) * W_{t-1}(\mathrm{s},\mathrm{a}) + \boxed{\mathrm{m}} * Q_{t-1}(\mathrm{s},\mathrm{a})\right) * \beta$$

- ➢ Q-actor-critic action values can then be used in a soft-max decision rule to calculate the probability of a given action

$$P(a_1,t)= e^{(Q\_AC(s,a_1,t))} / (e^{(Q\_AC(s,a_1,t))} + e^{(Q\_AC(s,a_2,t))}),$$

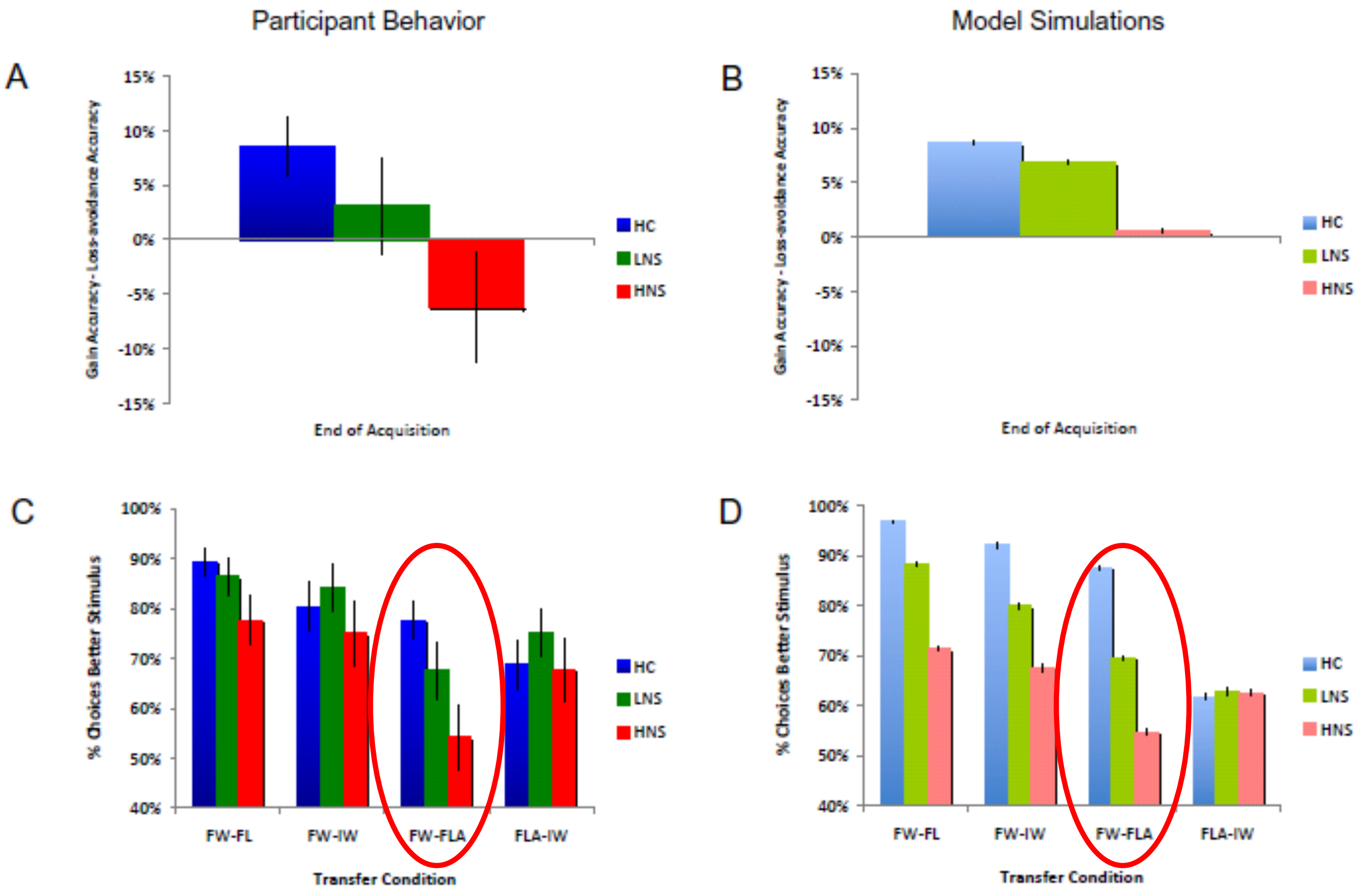
- ➢ The hybrid-probability model thus contains five free parameters: a *critic (α_C), actor (α_A), and Q (α_Q) learning rate, a temperature parameter (β) that captured how deterministically participants sampled the optimal choice, and a mixing (m) parameter that weighted the contributions of Q- and actor-critic-type learning.

Gain vs. Loss-avoidance Task (GLAT): Model Simulations



- Because the Q-learning model predicts sensitivity to actual outcome values, it predicts that subjects will choose a frequent winner over a frequent loss avoider.
- 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
- The AC model cannot account for the observed preference of frequent winners (FW) compared to frequent loss avoiders (FLA) across all groups, since both choices are likely to be associated with the same frequency of positive and negative prediction error, but frequent winners have higher expected outcome.

Gain vs. Loss-avoidance Task (GLAT): Model Simulations



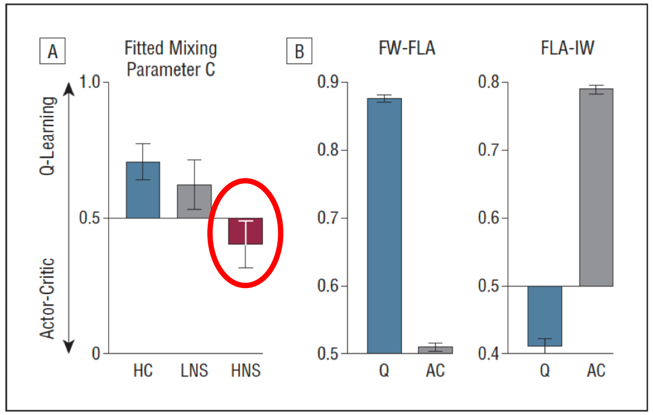


Figure 5. The relative contribution of Q learning and actor-critic learning to behavioral choices. A, Greater contribution of Q learning in healthy control (HC) subjects relative to the patient groups. Only the contrast between the HC group and the high-negative symptom (HNS) group was statistically 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.

Advantages of the Hybrid Model

- Because the Q-learning model predicts sensitivity to actual outcome values, it predicts that subjects will choose a frequent winner over a frequent loss avoider.
- 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
- The AC model cannot account for the observed preference of frequent winners (FW) compared to frequent loss avoiders (FLA) across all groups, since both choices are likely to be associated with the same frequency of positive and negative prediction error, but frequent winners have higher expected outcome.

Probabilistic Reinforcement Learning:
Behavioral Phenomena We Have Linked to Avolition/Anhedonia in SZ

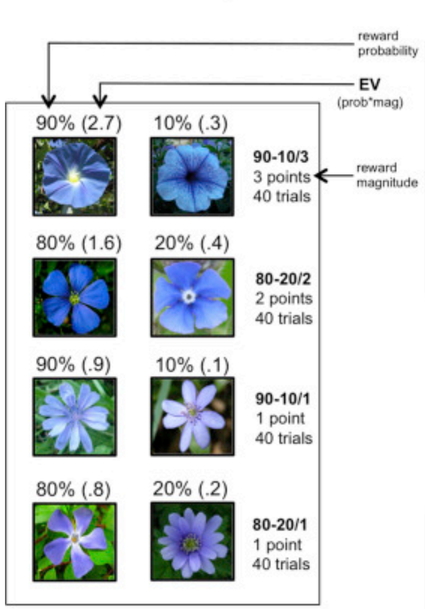
- Relatively-intact negative RPE-driven learning in the presence of impaired positive RPE-driven learning

- **Relatively-intact gradual/BG-driven/procedural learning in the presence of impaired rapid/OFC-driven/explicit RL**
    - Relatively-intact habit learning in the presence of impaired WM-dependent RL
    - Greater performance deficits in SZ for more deterministic contingencies than less deterministic contingencies (value-difference effect)
    - A reduced ability to integrate the frequencies and magnitudes of outcomes

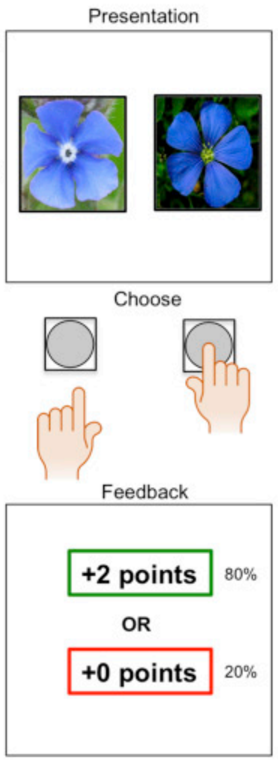
- In general: a more limited ability to rapidly and flexibly update value representations in the brain

Assessing the Ability to Integrate Reward Probability and Magnitude of Recent Outcomes with a Stimulus Selection Task

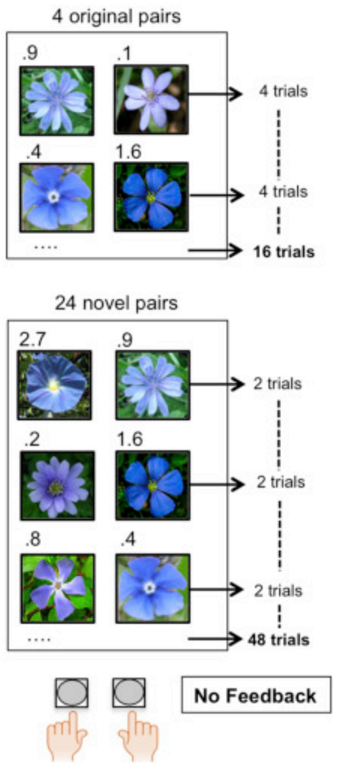
160-trial Learning Phase

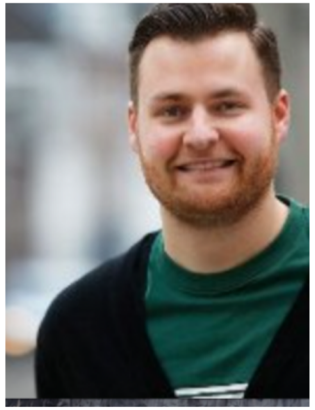


Trial Example



64-trial Transfer Phase



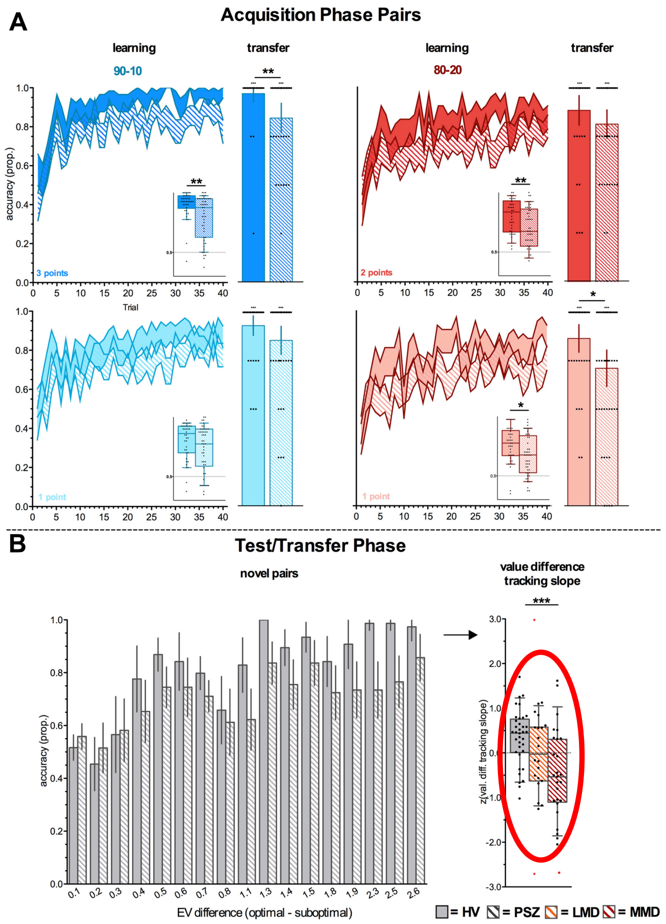


- **Learning Phase**
  - On each trial, two stimuli were presented, on either side of a fixation cross.
  - Participants were prompted to select one stimulus by pressing either the left or right trigger on a gamepad using their left or right index finger.
  - Each choice was followed immediately by feedback, in the form of a number of points (+3, +2, +1, or +0).
  - The eight stimuli differed in the probability and magnitude of the expected reward.
  - All pairs were presented 40 times in pseudorandomized order.
- **Test/Transfer Phase**
  - Purpose was to assess participants' ability to combine reward probability and magnitude into a representation of EV.
  - Participants were presented with the four familiar learning phase pairs ("acquisition pairs"; four presentations per pair) and 24 novel pairs of stimuli (two presentations per pair) and received the following instructions: "Please choose the picture that feels like it's worth more points based on what you have learned during the previous block."
  - Crucially, for many of these trials, the optimal answer depended on the ability to combine the expected probability and magnitude of a stimulus (e.g., 80/2 vs. 90/1, or 10/3 vs. 20/2).
  - No performance feedback was presented during the test/transfer phase.

From Hernaus et al. (2019). Impaired Expected Value Computations in Schizophrenia Are Associated With a Reduced Ability to Integrate Reward Probability and Magnitude of Recent Outcomes. *Biological Psychiatry: CNNI, 4,* 280-290.

Integrating Reward Probability and Magnitude: The “value difference tracking slope”

- HVs outperformed PSZ on all stimulus pairs
- Performance in block 4 (trials 31–40) was above chance in both participant groups for every pair
- There was also a main effect of pair, suggesting that both greater reward magnitude and probability conferred performance improvements.
- The **value difference tracking slope** was greater for HVs than PSZ
    - These data indicate that PSZ performance improved less as the difference in EV between two competing stimuli increased.
    - Importantly, the group difference in the value difference tracking slope was driven by motivational deficit (avolition/role-functioning and anhedonia/asociality subscales) severity (Panel B).
    - These results suggest that the MMD subgroup specifically was poorer at integrating reward probability and magnitude.

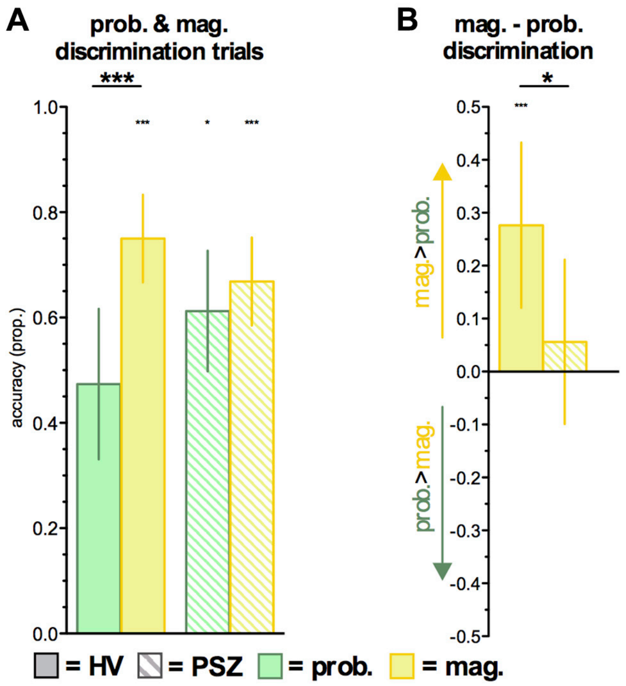


From Hernaus et al. (2019). Impaired Expected Value Computations in Schizophrenia Are Associated With a Reduced Ability to Integrate Reward Probability and Magnitude of Recent Outcomes. *Biological Psychiatry: CNNI, 4,* 280-290.

Controlling for Reward Probability and Magnitude

- On selective trials matched for probability and magnitude, HVs performed better on magnitude discrimination than probability discrimination trials, while PSZ performed similarly on magnitude and probability discrimination trials.

- The difference between performance on magnitude- and probability-discrimination trials—that is, the difference between the advantage conferred by higher reward magnitude versus higher reward probability—highly correlated with the value difference tracking slope, suggesting that participants who performed better on magnitude discrimination trials overall performed better in the test/transfer phase.



From Hernaus et al. (2019). Impaired Expected Value Computations in Schizophrenia Are Associated With a Reduced Ability to Integrate Reward Probability and Magnitude of Recent Outcomes. *Biological Psychiatry: CNNI, 4,* 280-290.

The Hybrid Model combines Actor-Critic and Q-learning mechanisms using a mixing parameter

$$(1)\ Q_t(\mathrm{s,a}) \ = \ Q_{t-1}(\mathrm{s,a}) + \alpha_Q * \delta(\mathrm{t})$$

$$(2)\ V_t(\mathrm{s}) = \ V_{t-1}(\mathrm{s}) + \alpha_C \ * \delta(\mathrm{t})$$

$$(3)\ W_t(\mathrm{s,a}) = \ W_{t-1}(\mathrm{s,a}) + \alpha_A \ \ * \delta(\mathrm{t})$$

$$(4)\ Q\_AC_t(\mathrm{s,a}) = \left((1 - \textcircled{m}) * W_{t-1}(\mathrm{s,a}) + \textcircled{\mathrm{m}} * Q_{t-1}(\mathrm{s,a})\right) * \beta$$

- Q-actor-critic action values can then be used in a soft-max decision rule to calculate the probability of a given action

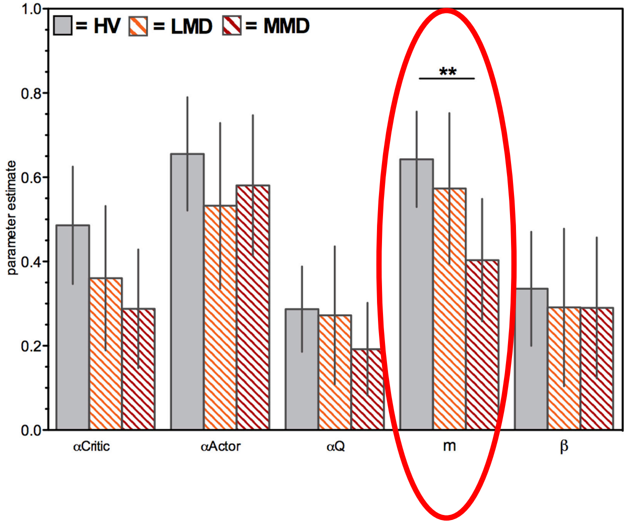
$$P(a_1,t) = e^{(Q\_AC(s,a_1,t))} / \left(e^{(Q\_AC(s,a_1,t))} + e^{(Q\_AC(s,a_2,t))}\right),$$

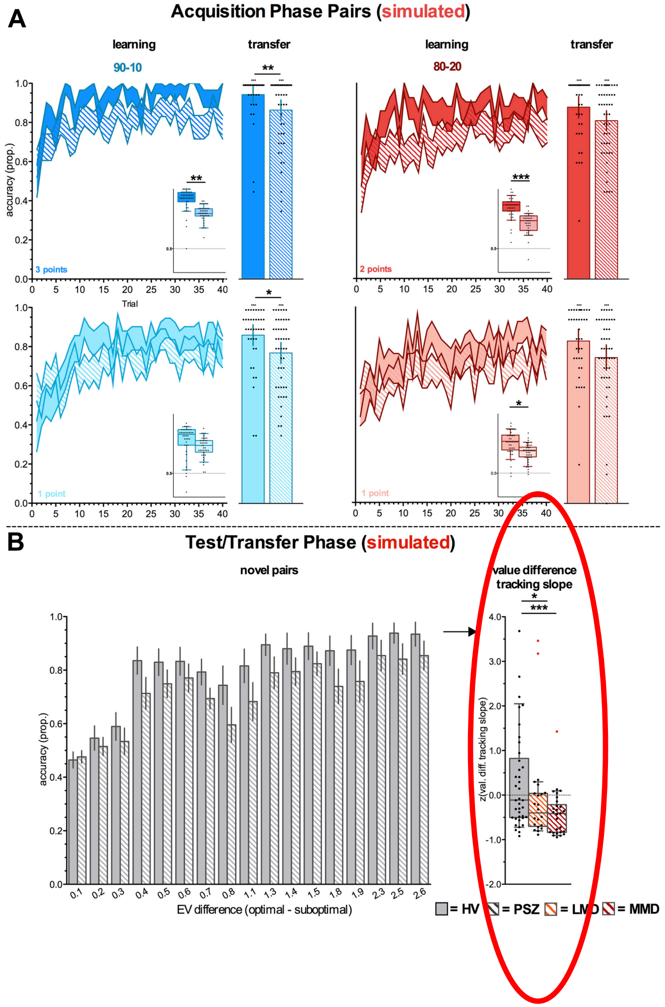
- The hybrid-probability model thus contains five free parameters: a *critic (α_C), actor (α_A), and Q (α_Q) learning rate, an inverse temperature parameter (β) that captured how deterministically participants sampled the optimal choice, and a mixing (m) parameter that weighted the contributions of Q- and actor-critic-type learning.

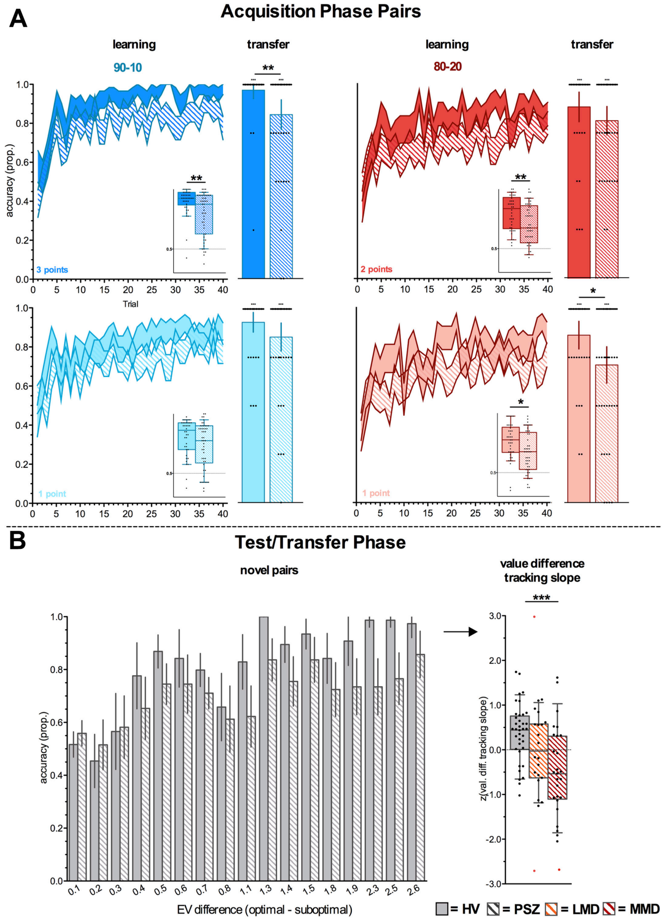
Modeling the Integration of Reward Probability and Magnitude

- Patients with the most severe motivational deficits showed the least contribution from the Q-learning component
- Reducing the Q-learning contribution had the effect of reducing the "value difference tracking slope"

hybrid-probability model parameters







From Hernaus et al. (2019). Impaired Expected Value Computations in Schizophrenia Are Associated With a Reduced Ability to Integrate Reward Probability and Magnitude of Recent Outcomes. *Biological Psychiatry: CNNI, 4,* 280-290.

Summary of Findings

- **In the context of an RL paradigm dependent on the successful integration of reward probability and magnitude**, PSZ – especially those with motivational deficits – were specifically impaired on trials with greater objective EV difference between two stimuli, as evidenced by the group difference in the test/transfer phase value difference tracking slope.

- Outcome probability-magnitude integration deficits in PSZ with motivational deficits were driven primarily by increased reliance on valueless stimulus-associations (actor-critic), in lieu of EV-based decision making (Q-learning).

- Individual value difference tracking slopes correlated significantly with estimates of individual mixing parameters, which capture the balance between Q-learning and actor-critic–type learning, suggesting a systematic relationship between EV-based learning and probability-magnitude integration.

- Individual value difference tracking slopes also correlated significantly with motivational deficit severity, thereby providing formal computational modeling evidence that impaired probability-magnitude integration, due to overutilization of stimulus-response associations, may play a role in the onset of motivational deficits PSZ.

Interpretations of Findings

- A reduced ability to combine reward magnitude and probability in the service of generating adaptive estimates of EV in PSZ with motivational deficits is in line with a large body of previous work, including findings of performance deficits in PSZ on the Iowa Gambling Task.

- Altogether, the current work reconfirms the notion that performance deficits in PSZ increase with demands placed on putative prefrontal processes involved in EV estimation.

- A failure to appropriately combine reward magnitude and probability into a single estimate of EV may lead to a decrease in perceived reward value, which may change the trade-off between reward and effort cost, and thus the willingness to exert effort, in line with findings that abnormal effort-cost computations are most pronounced in avolitional PSZ in conditions with high reward value.

Probabilistic Reinforcement Learning: Computational Explanations for Phenomena We Have Observed

- Attenuated learning rates for positive RPEs in schizophrenia, but relatively intact learning rates for negative RPEs
  - Relatively-intact negative RPE-driven learning in the presence of impaired positive RPE-driven learning
- Overreliance on Stimulus-Response (A/C) Learning in Schizophrenia, at the expense of Q-learning
  - Can account for relatively equal preference for Winners and Loss-avoiders among SZ with high avolition/anhedonia
  - Can account for a reduced ability to integrate reward probability and magnitude of recent outcomes in SZ

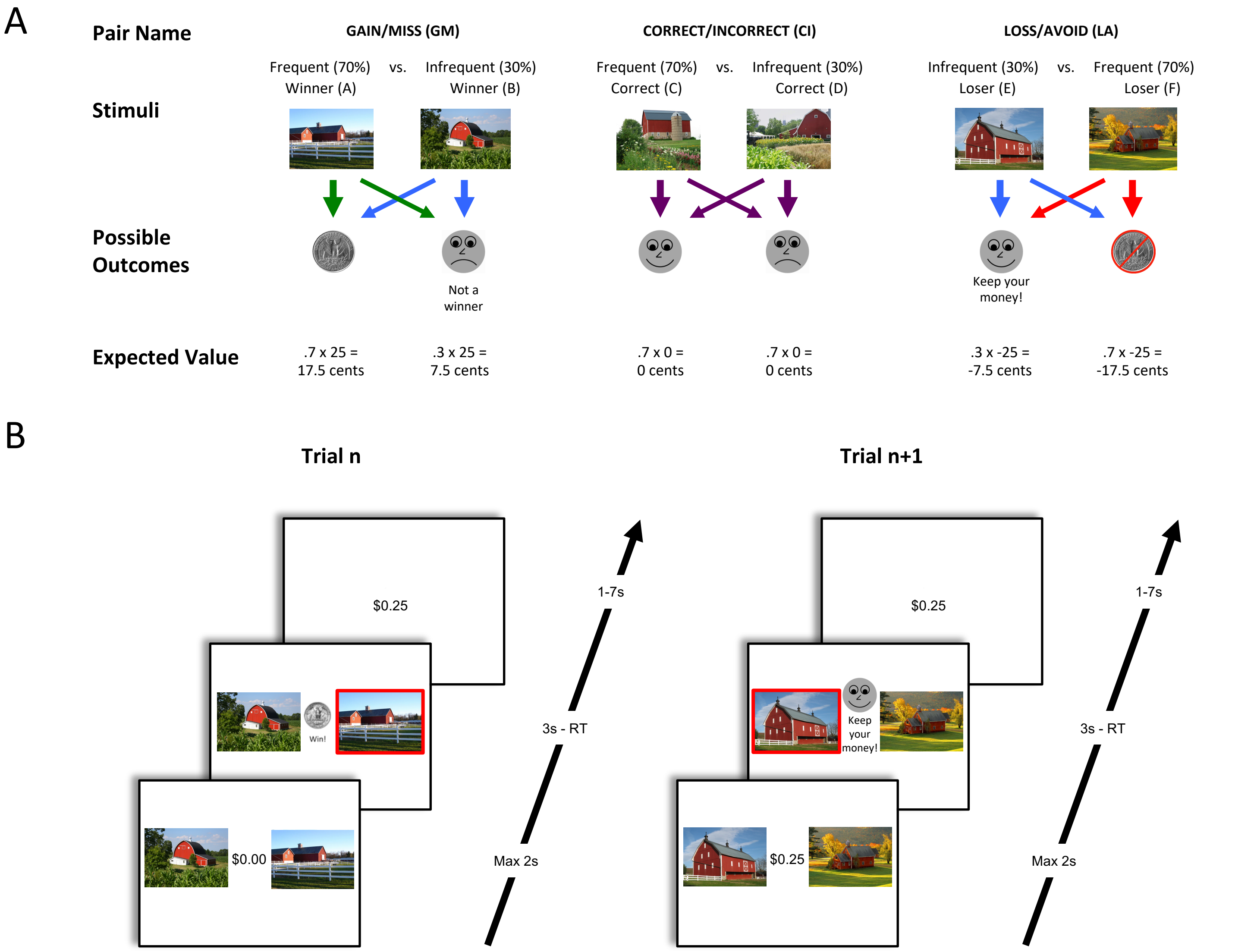
These are values/sets of values we might consider to be computational phenotypes in the context of our work…

Subjects can be characterized in terms of model parameters that correspond to constructs:

$\alpha_{G,}$ $\alpha_{N,}$ β → Trial-wise representations of EV and RPE

Using trial-wise parameters in neuroimaging studies of learning and decision making

Gain- vs. Loss-driven Learning:
The Probabilistic Stimulus Selection (PSS) Task



A

Pair Name	GAIN/MISS (GM)		CORRECT/INCORRECT (CI)		LOSS/AVOID (LA)	
Stimuli	Frequent (70%) Winner (A)	vs. Infrequent (30%) Winner (B)	Frequent (70%) Correct (C)	vs. Infrequent (30%) Correct (D)	Infrequent (30%) Loser (E)	vs. Frequent (70%) Loser (F)
Possible Outcomes		Not a winner			Keep your money!	
Expected Value	.7 x 25 = 17.5 cents	.3 x 25 = 7.5 cents	.7 x 0 = 0 cents	.7 x 0 = 0 cents	.3 x -25 = -7.5 cents	.7 x -25 = -17.5 cents

B

Trial n

Trial n+1

Two Learning Rate Q-learning Model

- Model value on a trial-wise basis, as $Q_i(t)$
- You do this by updating value as a function of the mismatch between the expected and obtained outcome at time t [r(t)]
  - This is the reward prediction error, called δ

$$\delta(t) = [r(t) - Q_i(t)]$$

- The actual change in value is a function of both δ and a parameter called Learning Rate (α), which is estimated for a group or individual
  - Determines the "impact" of prediction errors
  - We used separate learning rates for positive and negative RPEs (α_P and α_N):

$$\text{If } \delta \geq 0,\ Q_i(t+1) = Q_i(t) + \alpha_P \cdot \delta(t)$$

$$\text{If } \delta < 0,\ Q_i(t+1) = Q_i(t) + \alpha_N \cdot \delta(t)$$

  - Some frameworks call for modeling certainty about value on a trial-wise basis and use it to *estimate learning rate on a trial-wise basis*

2 Learning Rate Q-learning Model

- A decision function predicts the choice based on the relative values of the options:

$$P_i\ (t)\ = \frac{exp[\beta \cdot Q_i(t)]}{\sum_{k=1}^{n} exp[\beta \cdot Q_k(t)]}$$

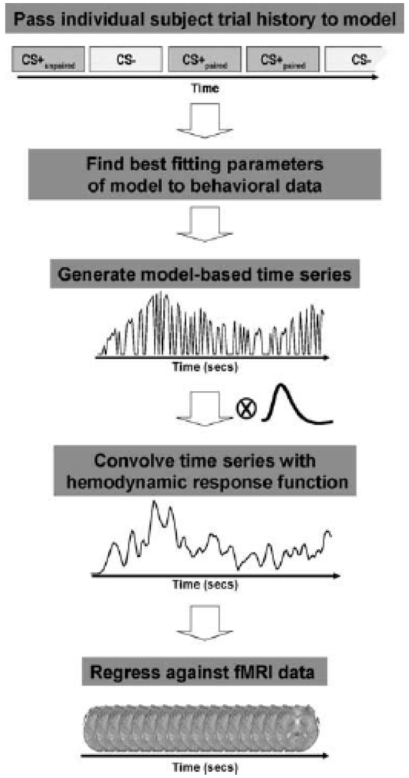
with a parameter beta (β) determining how strictly Q determined the choice of action

- The “fit” of the model is a function of how accurately it estimates individual choices and performance
  - Modeling scripts generate a value called the “log-likelihood estimate” (LLE), which is used to derived various measures of fit

- If the model fit is good enough, for an individual subject, these trial-wise model parameter estimates are what we use to create parametric regressors for fMRI data analysis.

Combining behavioral modeling with neuroimaging

1. Pass individual subject trial history to model
2. Find best-fitting parameters of model to behavioral data
3. Generate model-based time series
4. Convolve time series with hemodynamic response function
5. Regress against fMRI data



Pass individual subject trial history to model

Find best fitting parameters of model to behavioral data

Generate model-based time series

Convolve time series with hemodynamic response function

Regress against fMRI data

FIGURE 1. Illustration of model-based fMRI approach. Each individual subject's trial history is passed to the model, and the parameters of the model are fit so as to minimize the difference between the model predictions and an external behavioral measure, which in the conditioning example could be an external measure of conditioning, such as galvanic skin conductance responses or pupil dilation. Next, the best model-fitting parameters are used to generate a time series for each trial in the fMRI, which are then convolved with basis function(s) to account for the effects of hemodynamic lag, such as the canonical hemodynamic response function, and then regressed against the fMRI data.

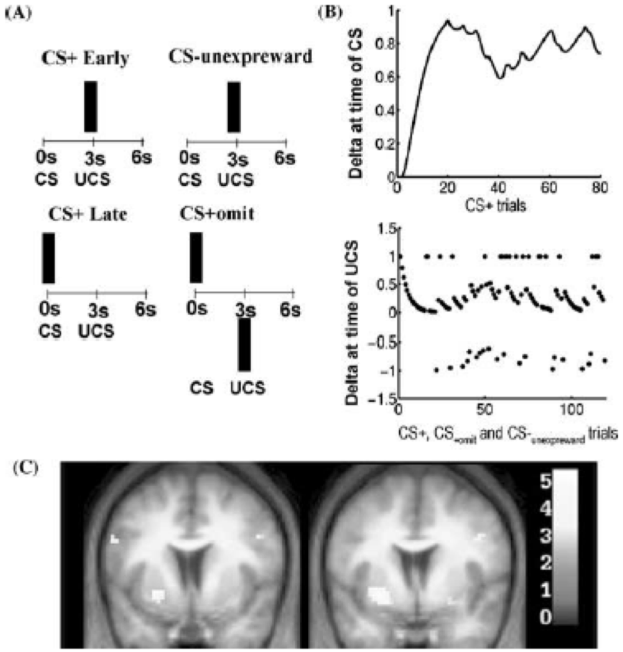
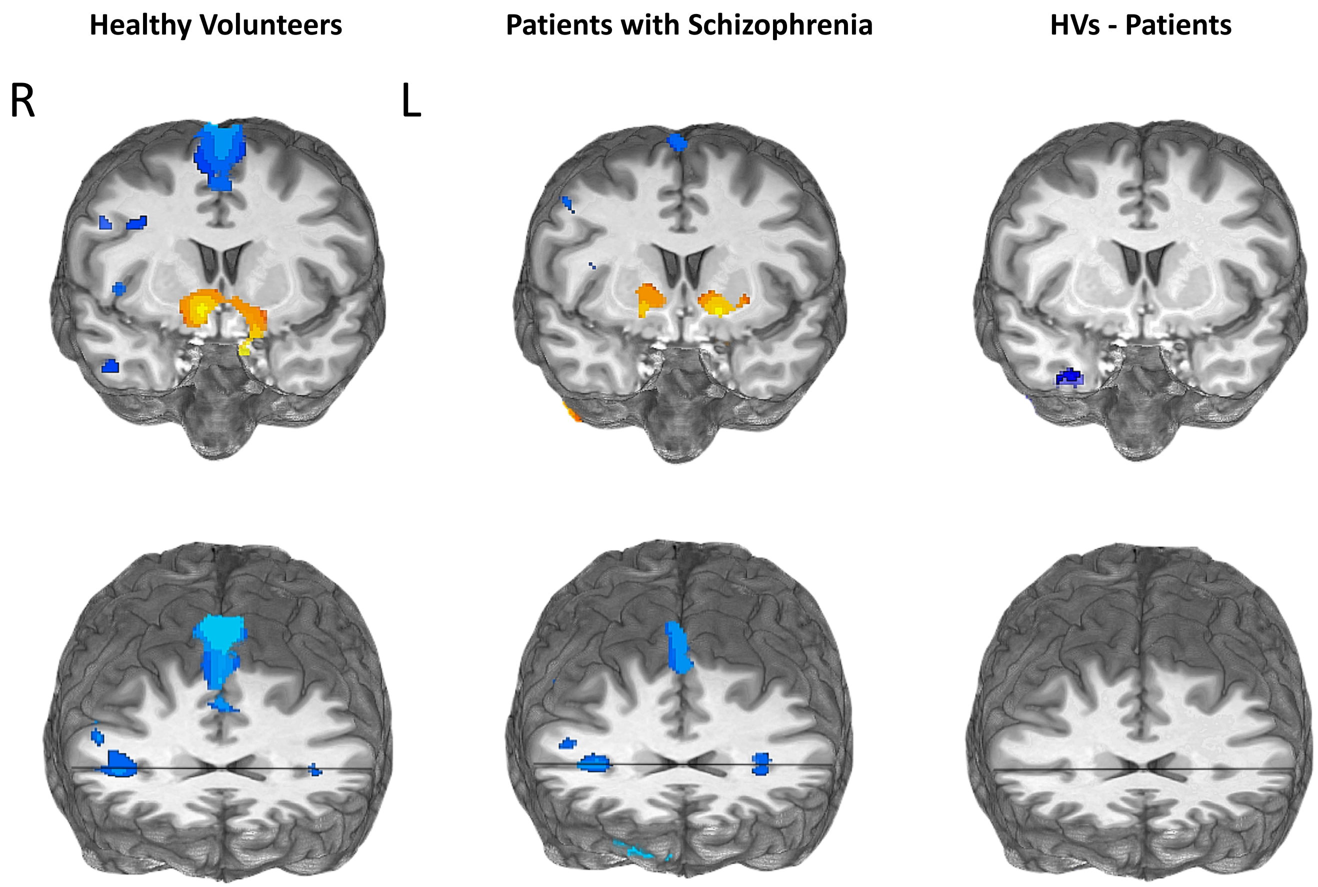


FIGURE 2. Model-based fMRI of stimulus-reward learning. (**A**) Properties of the temporal difference prediction error signal during reward learning in which a cue (CS+) is paired repeatedly with a reward (UCS) presented 3 sec later. During the initial stages of learning (CS + early trials), the error signal responds at the time of presentation of the UCS, but over the course of learning transfers back to the time of presentation of the CS (CS + late trials). On trials in which the CS+ is not presented but the reward is delivered anyway (CS–unexp. reward), the signal shows a positive response at the time the reward is delivered, whereas on trials in which the CS is presented but the reward is unexpectedly omitted the signals show a negative response at the time of outcome. (**B**) Plot of model-generated prediction error signals at the time of presentation of the CS, and the time of presentation of the UCS, over the course of the experiment for a typical subject. (**C**) Area of bilateral ventral striatum (ventral putamen bilaterally) showing significant correlations with the temporal difference prediction error signal while subjects underwent classical conditioning with sweet taste reward (1M glucose). Data from O'Doherty *et al.*[11]

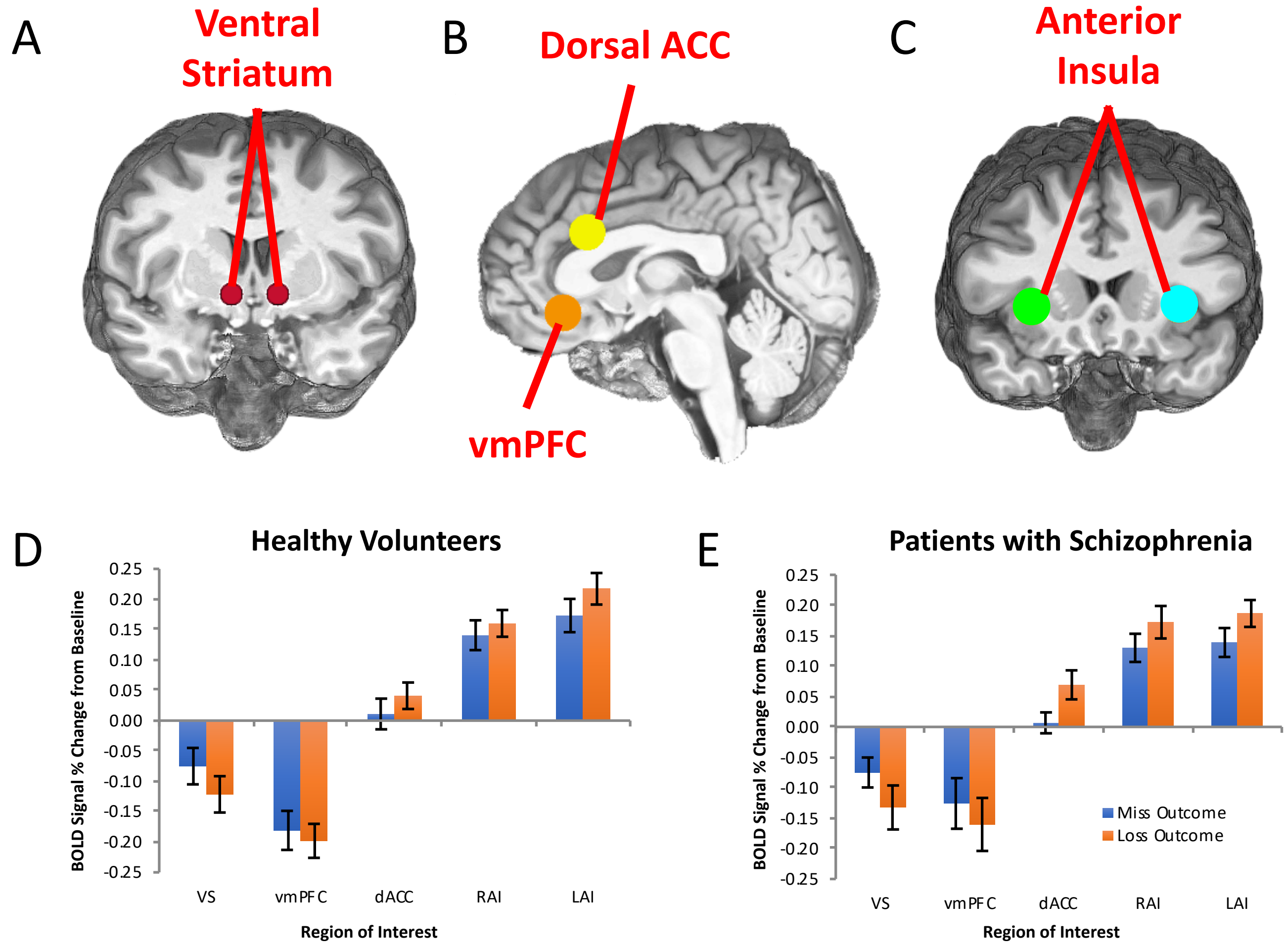
O'Doherty, J.P., et al. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. *ANYAS, 1104,* 35–53.

SZ and controls show similar neural responses to RPEs, in striatum, insula, and dmPFC

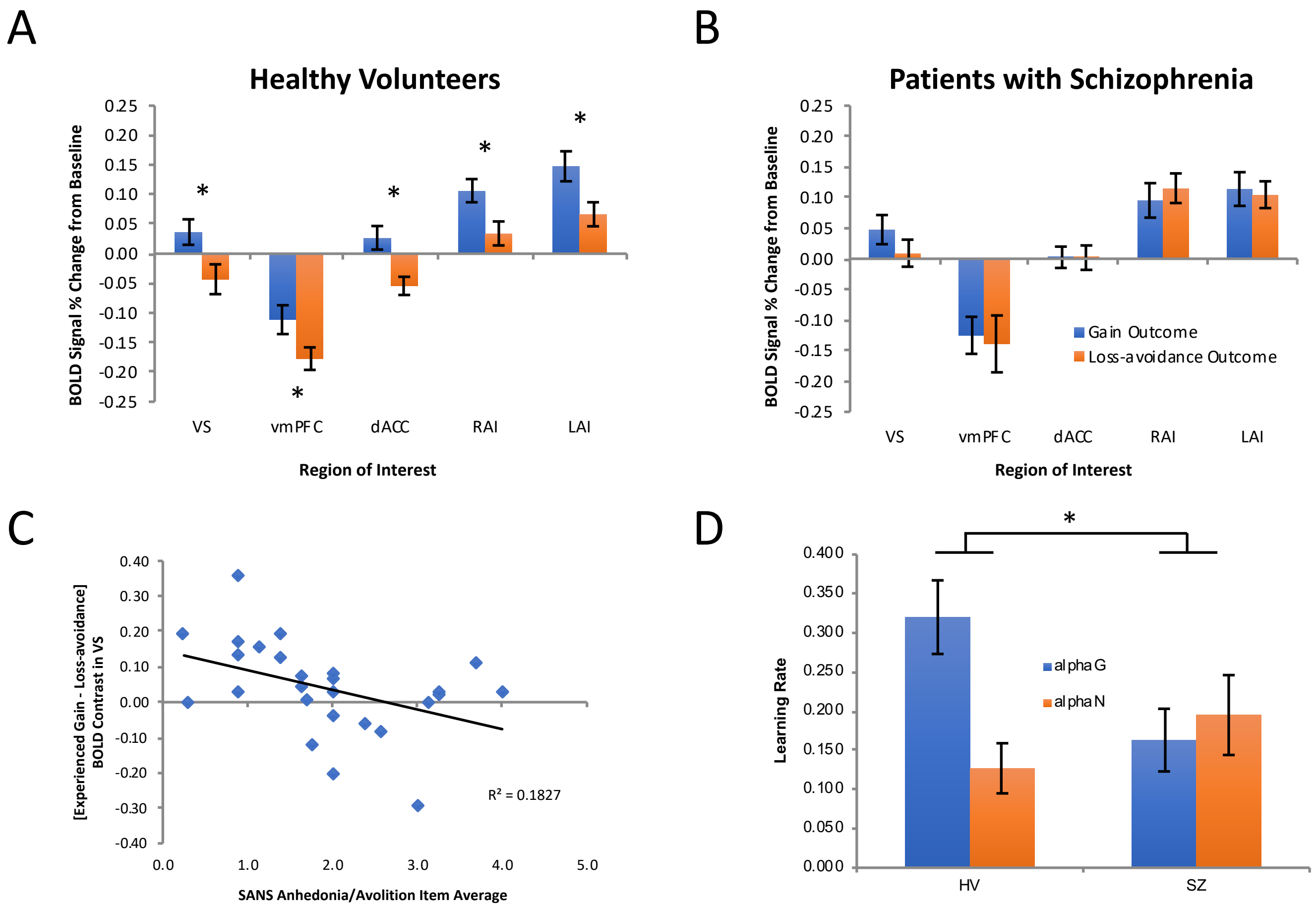


Waltz et al. (2018). BPS: CNNI, 3, 239.

Both patients and controls show similar neural responses to misses and losses (both negative RPEs)



Waltz et al. (In Press). BPS: CNNI.



- ➢ Controls show differential neural responses to gains and instances of loss-avoidance, but patients do not
- ➢ Experienced value [Gain – Loss-avoidance] contrasts in VS correlate with ratings for avolition/anhedonia in SZs.
- ➢ Controls show greater learning from gains positive than negative RPEs, but patients do not.

Waltz et al. (2018). BPS: CNNI.

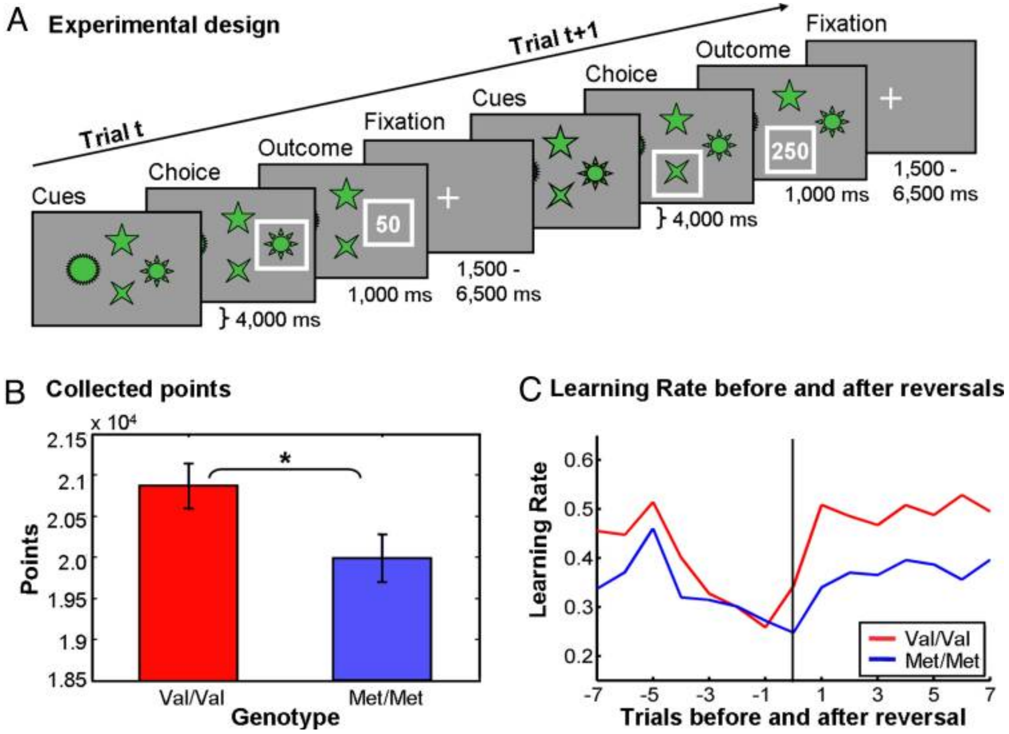
Summary of Computational Neuroimaging Findings from Gain- vs. Loss-driven PSS Task

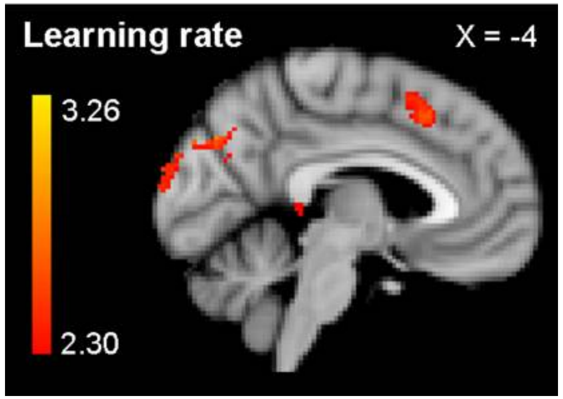
- Negative symptom scores in SZ patients correlated significantly with neural activity related to expected value-related activity in VS and vmPFC

- Suggests a specific deficit in representing the value of gains, in medicated SZ patients, with value updating being disproportionately influenced by learning about potentially negative consequences, as opposed to potentially positive ones

IV. Modeling Probabilistic RL in an Unstable Environment

Probabilistic RL in an Unstable Environment: COMT Genotype and Cognitive Flexibility



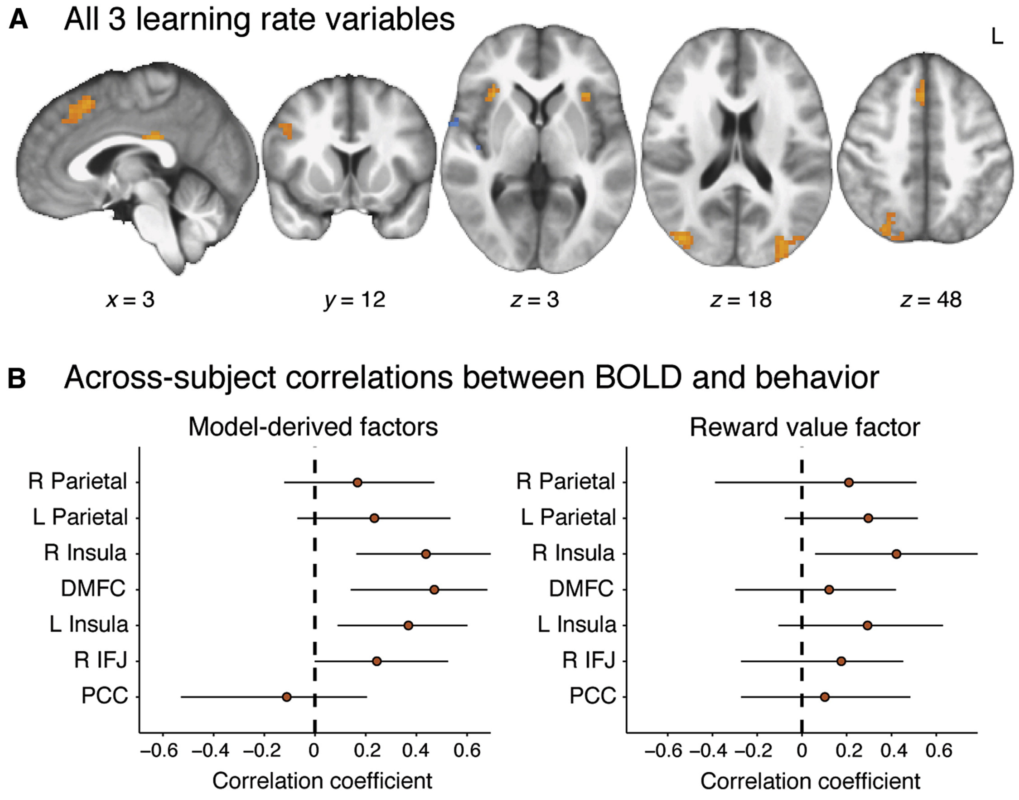


- Learning rate modeled as dynamic, varying with uncertainty (RPE slope)
    - Decreased with repeated positive feedback, increased with surprising negative feedback
- When used as a trial-wise parametric regressor in fMRI analyses, tracked by activity in dorsomedial frontal cortex (DMFC)

Krugel et al. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *PNAS, 106,* 17951-6.

Neural Substrates of Adaptive Learning

- Adaptive learning can be decomposed into three computationally and neuroanatomically distinct factors that were evident in human subjects performing a spatial-prediction task:
  1) surprise-driven belief updating, related to BOLD activity in visual cortex;
  2) uncertainty-driven belief updating, related to anterior prefrontal and parietal activity; and
  3) reward-driven belief updating, a context-inappropriate behavioral tendency related to activity in ventral striatum.
- These distinct factors converged in a core system centered on dorsomedial frontal cortex



A All 3 learning rate variables

B Across-subject correlations between BOLD and behavior

McGuire et al. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron, 84,* 870-881.

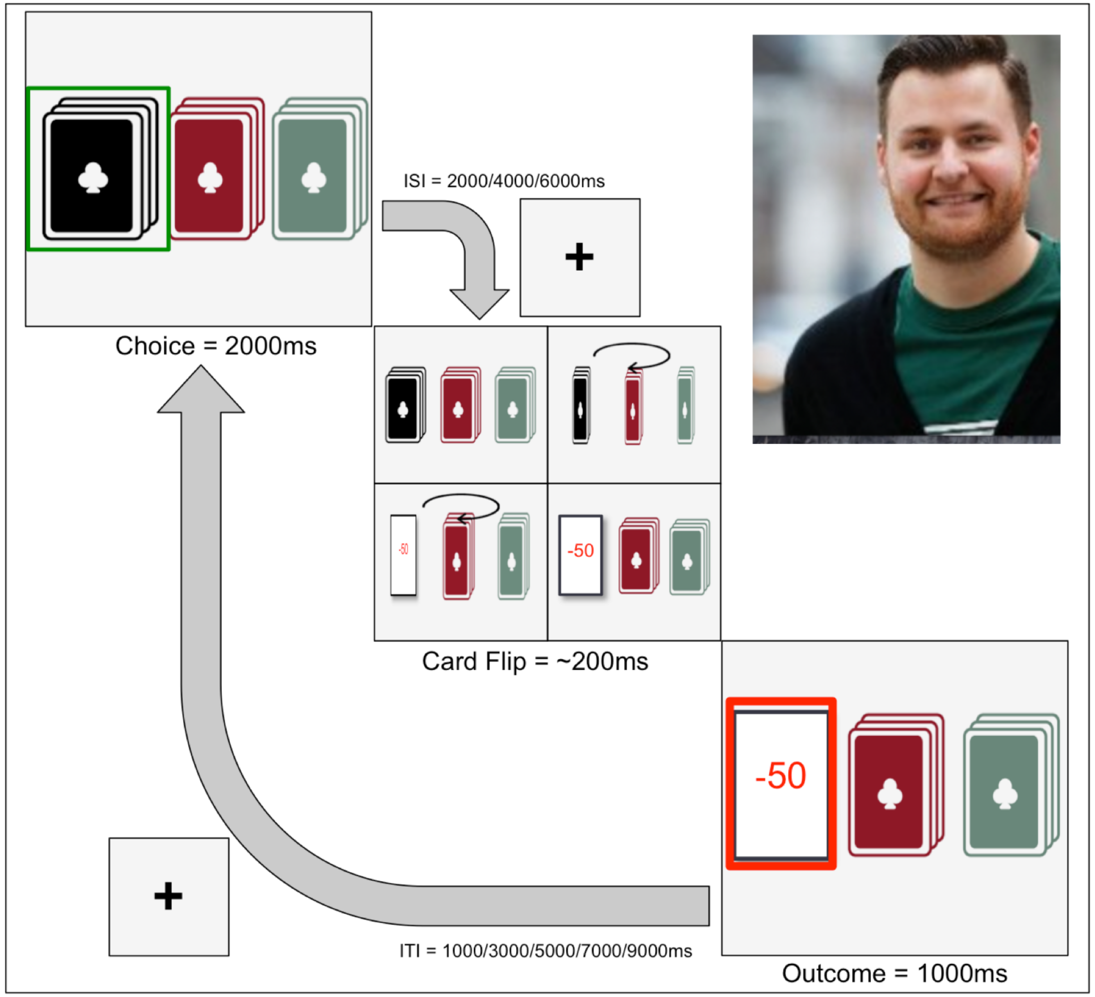
Hypotheses

- One way for PSZ to have reinforcement learning deficits in the presence of intact RPE signaling, is have a reduced ability to adaptive modulate learning rate as a function of uncertainty.

- A reduced ability to adaptive modulate learning rate as a function of uncertainty would be associated with abnormal activation of DMFC and/or connectivity between DMFC and ventral striatum.

The Card Betting RL Task

- The RL task consisted of a choice, card-flip and outcome phase.
- Choices were rewarded probabilistically, with a choice of the "optimal deck" leading to a 100-point gain on 90% of trials (and a loss of 50 points on 10% of trials). Choices of two non-optimal decks led to 100-point gains on 50% and 10% of trials (and losses of 50 points on 50% and 90% of trials), respectively.
- Participants were instructed to try to identify the optimal deck (i.e., the one with the highest expected value) as quickly as possible; they were also informed that, occasionally, a new deck would become the optimal one.
- Inter-trial (ITI) and inter-stimulus (ISI) intervals were pseudo-randomized, such that the average trial-length was ca. 10.5 s.
- Participants achieved as many stages as possible in 160 total trials (subdivided into 4 runs of 40 trials)



Modeling a Dynamic Learning Rate

- Participants' choices, and the outcomes of those choices, were fit to an RL model with a dynamic learning rate (α), such that EV, RPE, and α could be estimated on a trial-wise basis, based on the work of Krugel et al. (2009).

- In this model, learning rate on a given trial was updated as a function of m, the slope of the absolute values of consecutive RPEs.

- The impact of the RPE slope on α updating is determined by a parameter, β, which was optimized to fit individual subject data:

$$f(m)=\mathrm{sgn}(m)\cdot[1-e-(m/\beta)2].$$

- For increasing RPEs (m>0), learning rate updating would proceed as follows:

$$\boxed{\alpha(t)=\alpha(t-1)+f[(m)t]\cdot[1-\alpha(t-1)].}$$

- On any given trial, then, Q-values are updated as a function of the RPE and the learning rate at time i:

$$q_i(t)=q_i(t-1)+\alpha_i\cdot\delta_i(t-1).$$

Analyses of Model Parameters and Associated Neural Signals

- In order to compute group-level statistics, we extracted individual β parameters (signifying the impact of the RPE slope on α updating), as well as trial-wise learning rate estimates 8 trials before and after every reward contingency shift, for every individual.

- We then computed the slopes of learning rates prior to, surrounding, and following each reward contingency shift for a given individual.

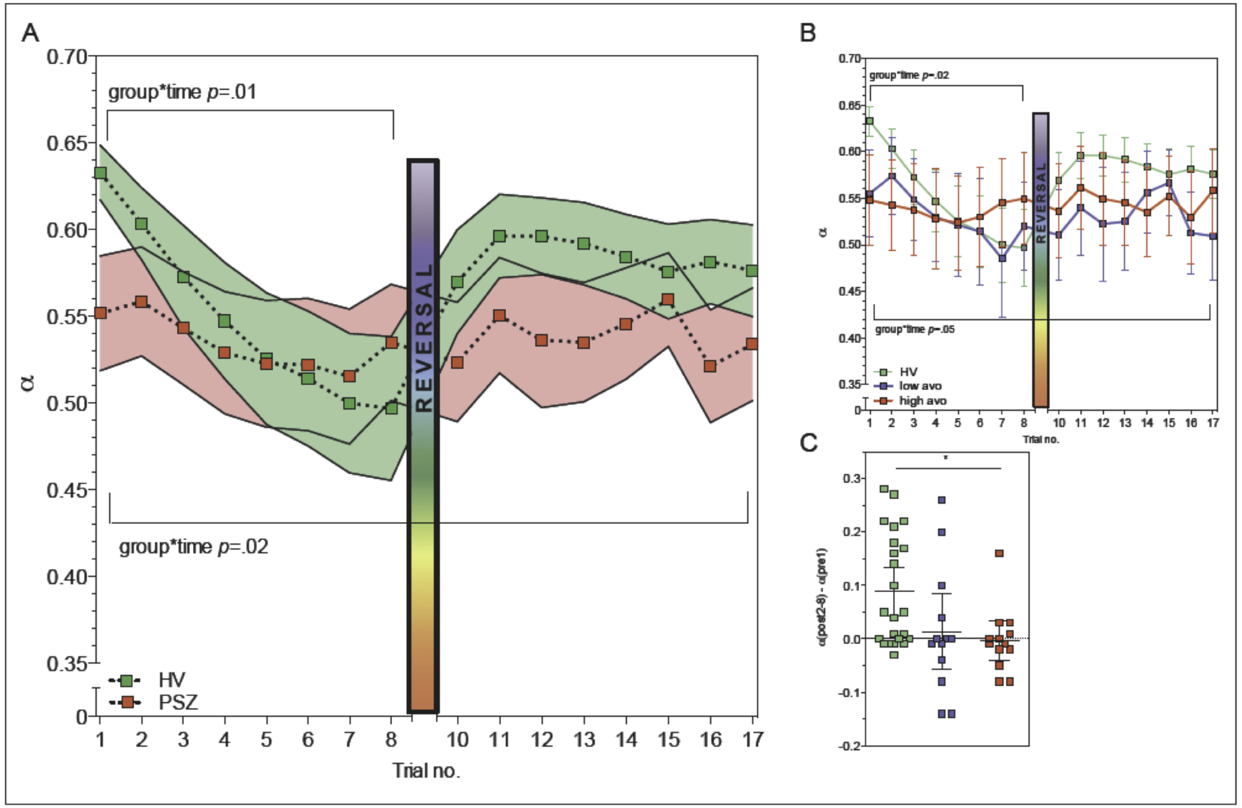
Learning rate modulation deficits increase with motivational deficit severity

A. Trial-wise estimates of learning rate 8 trials before and after a reinforcement contingency shift. PSZ, relative to HV, demonstrated a decrease in learning rate modulation, especially so in trials leading up to a contingency shift (Figure 2A).

B. This was especially the case for PSZ with high motivational deficits (Figure 2B), who showed little to no learning rate modulation across all trials.

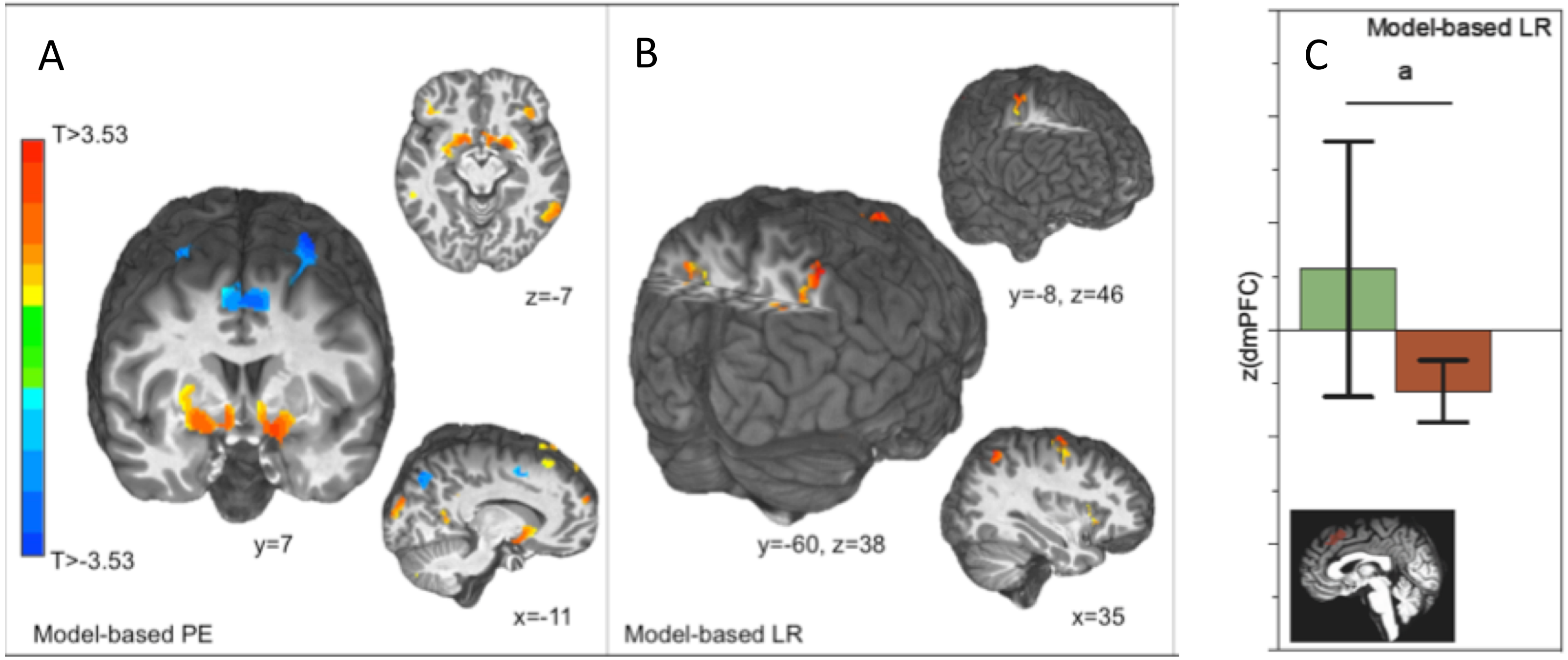
C. PSZ additionally demonstrated decreased post-shift impairments in learning rate modulation, defined as the difference between trial 1 pre-shift and trial 2-8 post-shift.

Panel A solid bars represent SEM. *=p<.05



Hernaus et al. (2018). Motivational deficits in schizophrenia relate to abnormalities in cortical learning rate signals. *Cogn Affect Behav Neurosci, 18,* 1338-1351

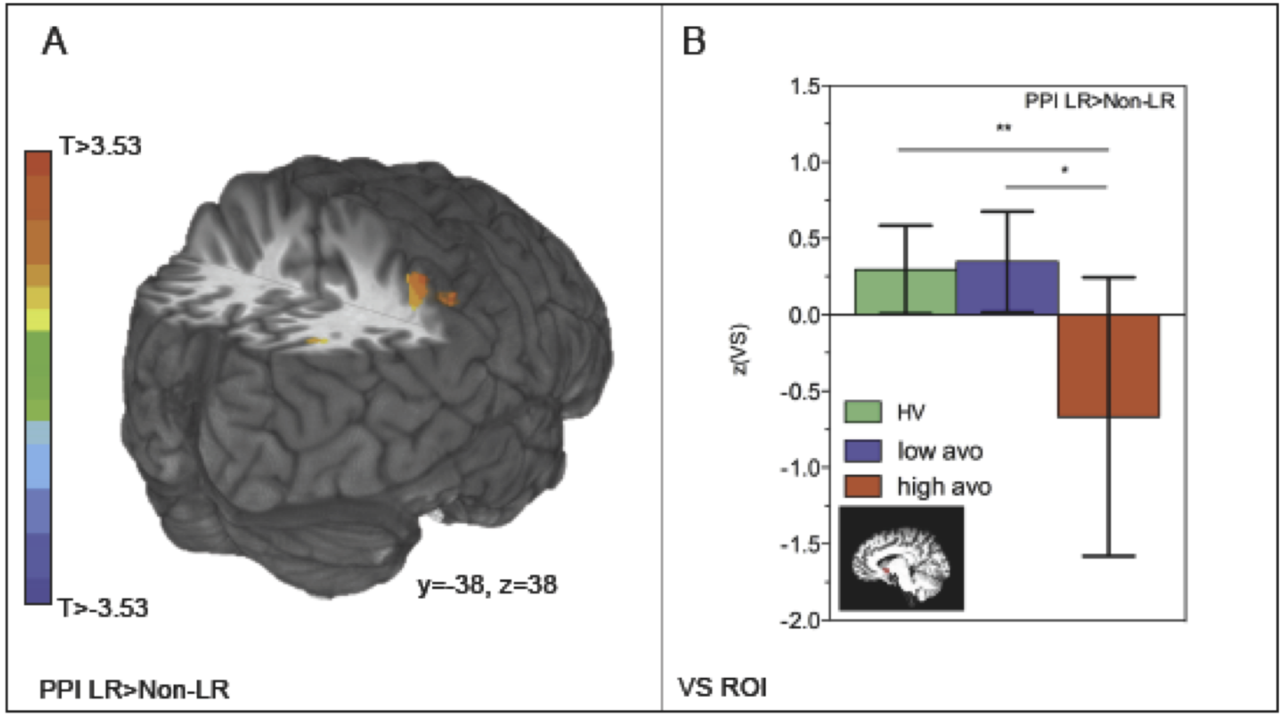
Model-based fMRI analyses, using regressors constructed from learning parameters



A. Robust RPE signals were observed in VS and related regions, while B. robust LR signals were observed in superior parietal lobule and dorsomedial prefrontal cortex (dmPFC). C. A trend toward a between-group difference in dmPFC was observed for the model-based learning rate analysis.

Hernaus et al. (2018). Motivational deficits in schizophrenia relate to abnormalities in cortical learning rate signals. *Cogn Affect Behav Neurosci, 18,* 1338-1351

Decreased dmPFC-VS coupling in SZ patients with high motivational deficits



A. Whole-brain functional connectivity between dmPFC and inferior parietal lobule increased in the entire sample from non-learning rate to learning rate trials. B. In a follow-up ROI analysis in VS, functional connectivity increases were observed from non-learning to learning-rate trials for HV and PSZ with low motivational deficits, while dmPFC-VS connectivity decreases were observed in PSZ with high motivational deficits (5B). **, p<.01; *, p<.05, bars represent 95% confidence intervals. LR=learning rate.

Hernaus et al. (2018). Motivational deficits in schizophrenia relate to abnormalities in cortical learning rate signals. *Cogn Affect Behav Neurosci, 18,* 1338-1351

Uncertainty does not only influence the rate at which one learns; it also influences one's drive to seek information

The Explore/Exploit Trade-off

V. Modeling Goal-directed Exploration

The Temporal Utility Integration Task (TUIT; Time Conflict Task)

Ready?

0

Ready?

78

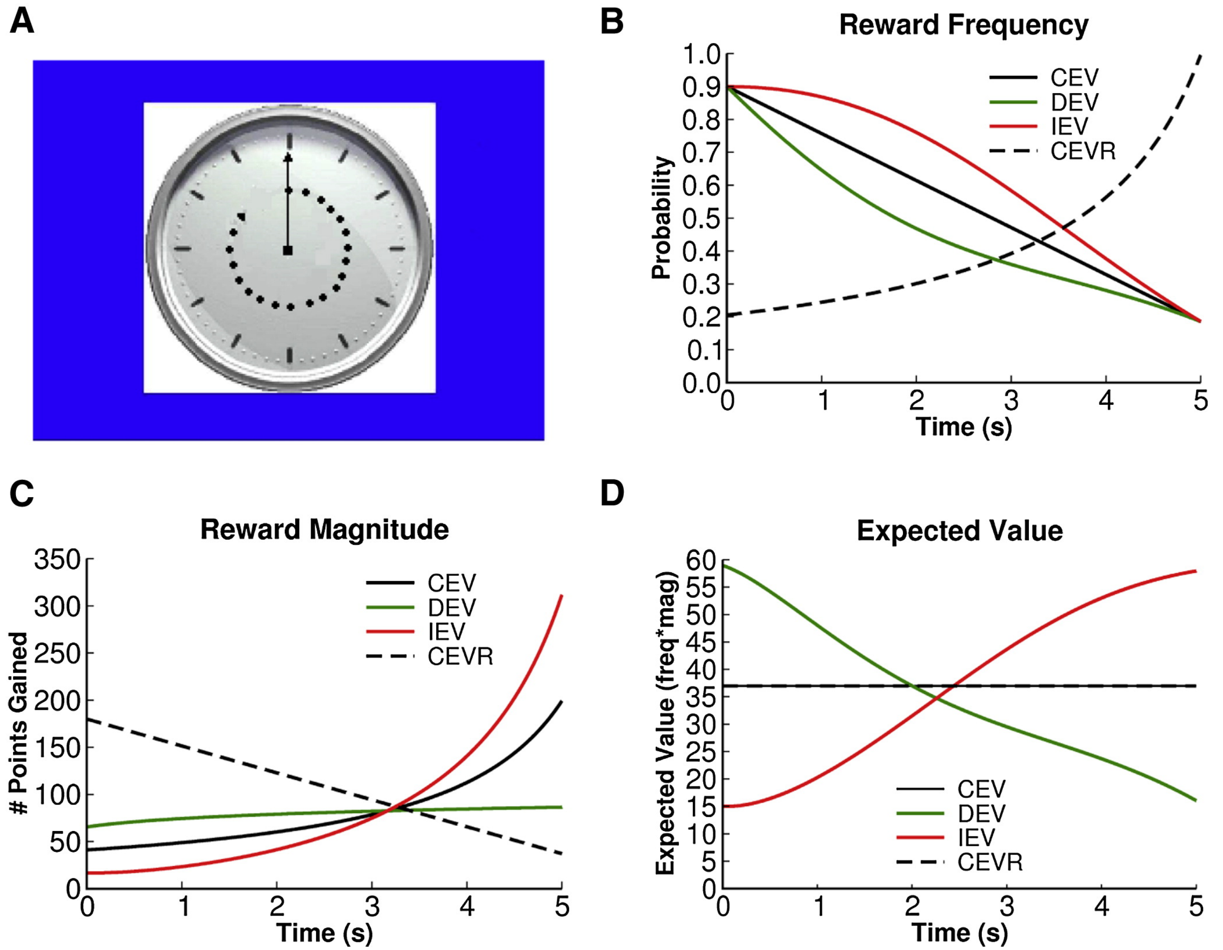
Ready?

182

Ready?

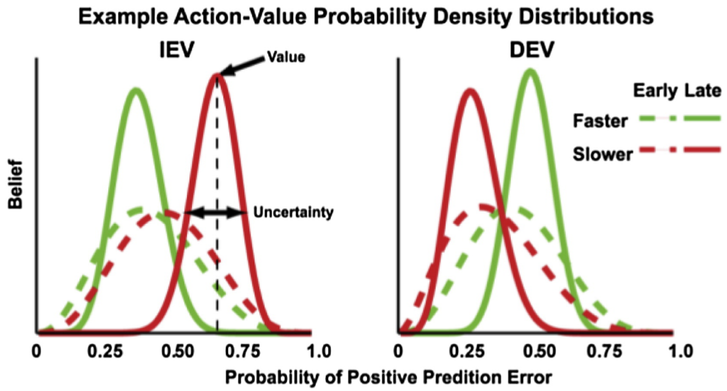
0

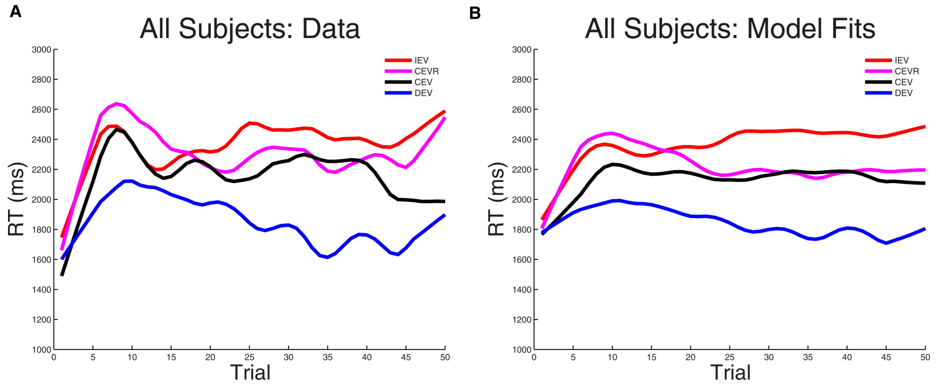
Temporal Utility Integration (TUI) task



From Moustafa et al. (2008). JNS, 28, 12294-11304.

Modeling behavior on the Temporal Utility Integration (TUI) task





From Badre et al. (2012). Neuron, 73, 595–607.

- We used a modeling approach based on the assumption that participants track the EV for the reward they expect to gain in a given block of trials
- When a given reward is greater or less than this EV, the associated prediction error signals drive learning to adjust behavior in two ways.
- First, a simple, likely implicit, process whereby accumulated positive RPEs translate into approach-related speeded responses (Go learning), whereas accumulated negative PEs produce relative avoidance and slowed responses (NoGo learning);

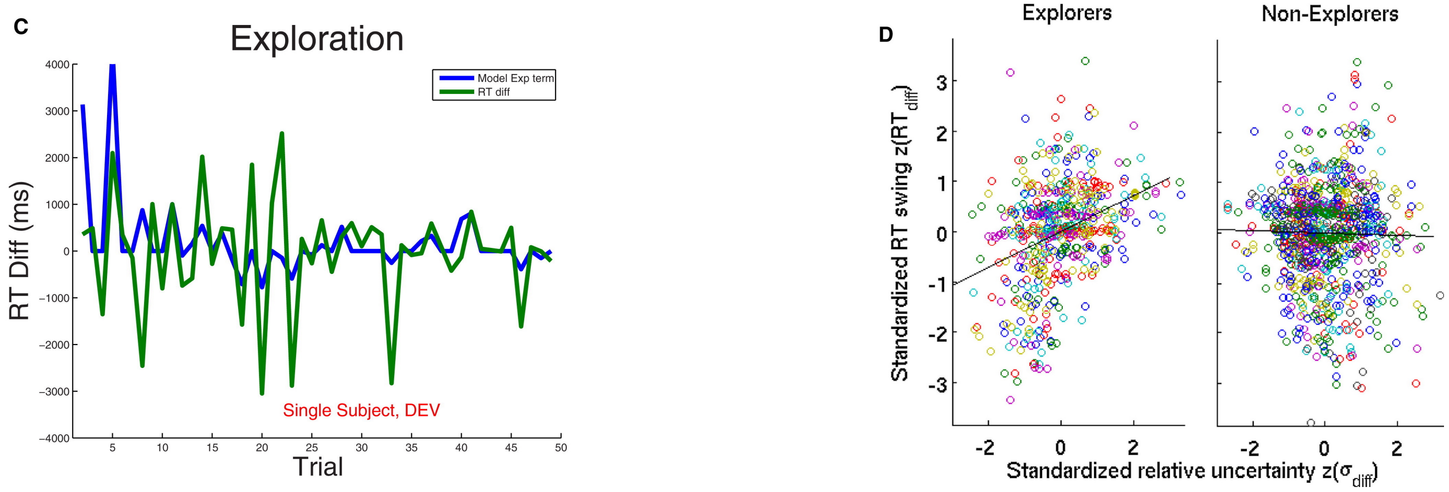
$$\mathrm{Go}(t) = \mathrm{Go}(t\text{-}1) + \alpha_{\mathrm{G}}\delta_{+}(t\text{-}1)$$
$$\mathrm{NoGo}(t) = \mathrm{NoGo}(t\text{-}1) + \alpha_{\mathrm{N}}\delta_{-}(t\text{-}1)$$

- The RL model of this task also assumed that individuals also use the relative difference in uncertainties about values to drive exploratory RT swings.
- Model thus incudes an exploration parameter, ε, predicting trial-by-trial RT swings to occur when one is relatively more uncertain about probability of obtaining a positive outcome for fast or slow responses.

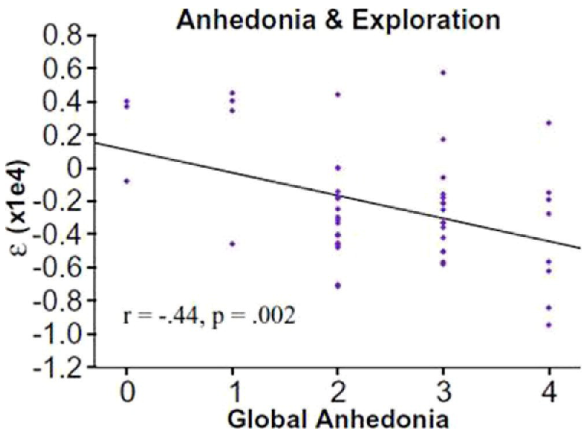
$$\mathrm{Explore}(t) = \varepsilon[\sigma_{\mathrm{slow}}(t) - \sigma_{\mathrm{fast}}(t)],$$

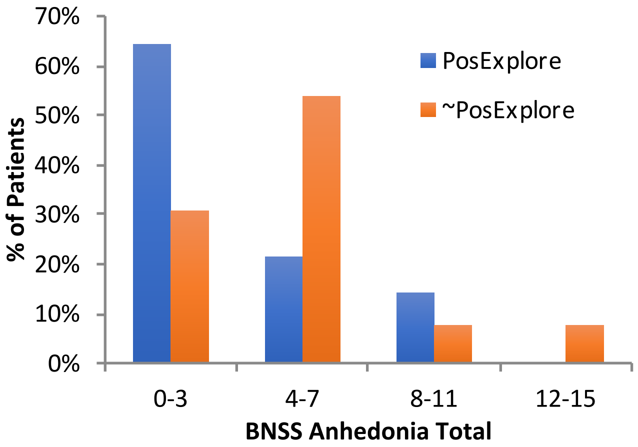
$$\hat{RT}(t) = K + \lambda RT(t-1) - Go(t) + NoGo(t) +$$
$$\rho[\mu_{slow}(t) - \mu_{fast}(t)] + \nu[RT_{best} - RT_{avg}] + Explore(t)$$

Investigating uncertainty-driven exploration in SZ using the Temporal Utility Integration (TUI) task



From Badre et al. (2012). Neuron, 73, 595–607.

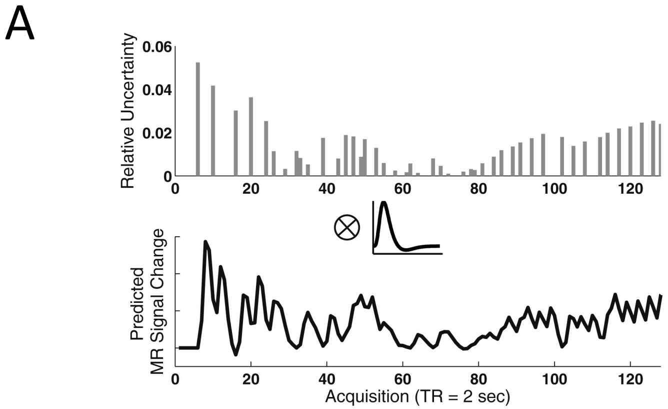




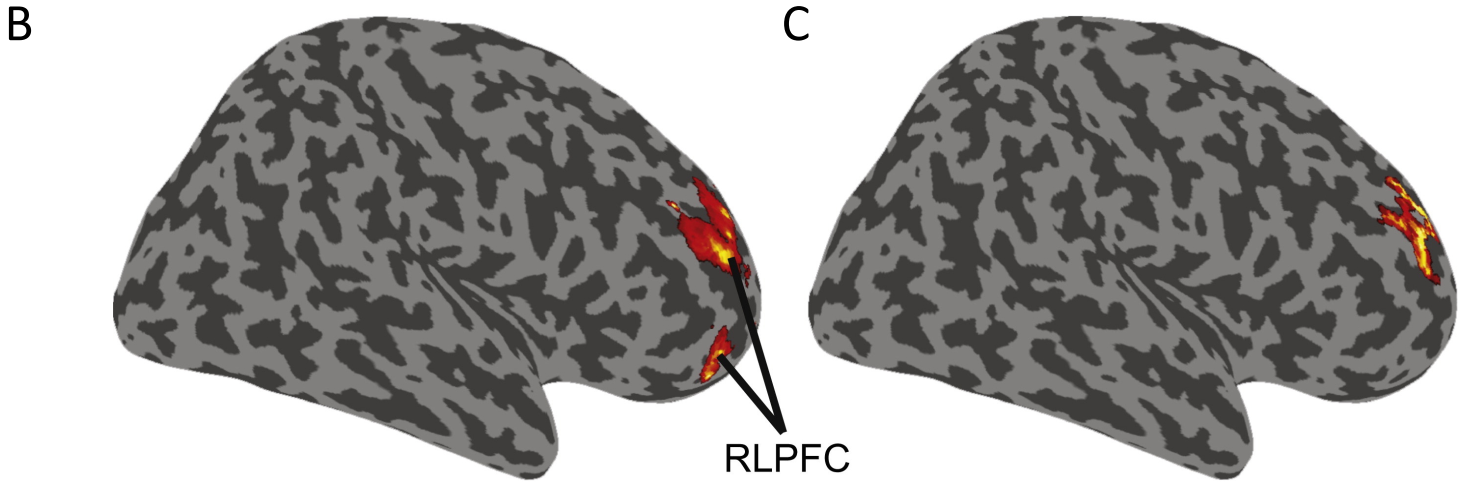
Strauss et al. (2011), Biol. Psychiat., 69, 424-431.

Waltz et al. (unpublished)

Neural Signature of Goal-directed Exploration



Whole brain analysis of trial-to-trial changes in relative uncertainty. (a) Example individual subject relative uncertainty regressor from one run of one participant. Convolution of parametric changes in relative uncertainty ($(|\sigma_{slow}(t) - \sigma_{fast}(t)|)$) on each trial (top plot) with a canonical hemodynamic response function (middle plot) produced individual participant relative uncertainty regressors (bottom plot).

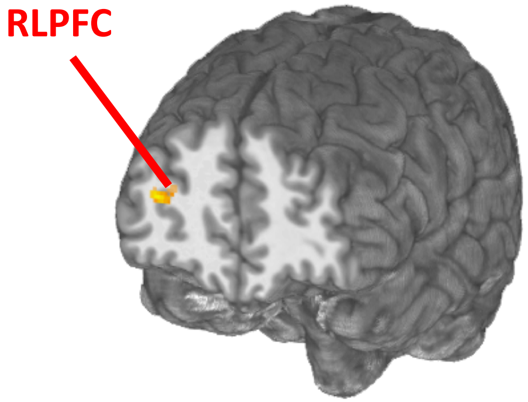


Explore Participants Only

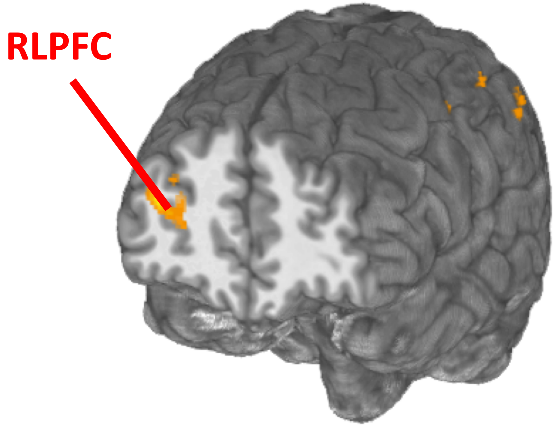
Explore – Nonexplore Participants

From Badre et al. (2012). Neuron, 73, 595–607.

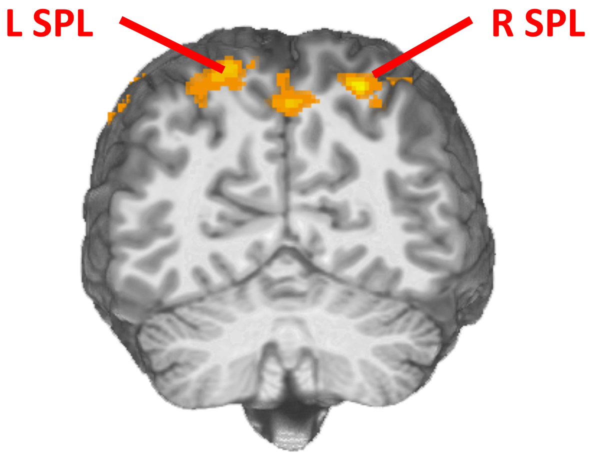
Schizophrenia and the Neural Substrates of Goal-directed Exploration



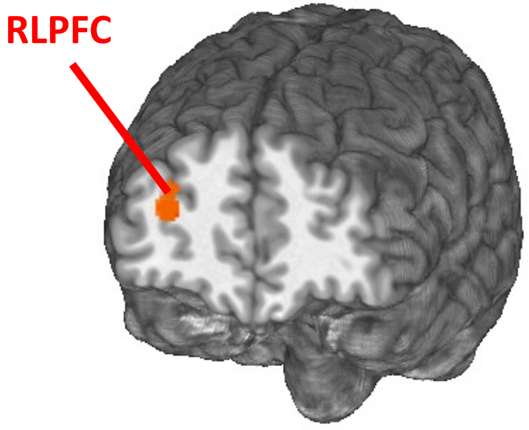
All Explore > All Nonexplore



All NCs > All SZs



All NCs > All SZs



NC Explore > SZ Explore

Waltz et al. (unpublished)

Uncertainty-driven Exploration in SZ: Conclusions and Interpretations

- These results suggest that motivational deficits in schizophrenia are associated with a specific type of goal-directed behavior: the tendency to explore reward contingencies based on uncertainty.

- Furthermore, people with schizophrenia show aberrant activity in the neural circuitry associated with the performance of uncertainty-driven exploration.

- These findings suggest an alternate source of motivation deficits in schizophrenia, aside from a reduced tendency to exploit known reward contingencies.

- In addition they link a specific deficit to abnormal activity in a specific circuit, which might serve as a biomarker for future studies of candidate risk genes or potential interventions.

VI. Lessons Learned

Computational accounts of motivational deficits in psychotic illness

- Relatively-intact negative RPE-driven learning in the presence of impaired positive RPE-driven learning
  - Attenuated learning rates for positive RPEs in schizophrenia, but relatively intact learning rates for negative RPEs

- Relatively equal preference for Winners and Loss-avoiders among SZ with high avolition/anhedonia
  - Can be attributed to overreliance on Stimulus-Response (A/C) Learning in Schizophrenia, at the expense of Q-learning (reduced mixing parameter, m)

- A reduced ability to integrate reward probability and magnitude of recent outcomes in SZ
  - Can also be attributed to overreliance on Stimulus-Response (A/C) Learning in Schizophrenia, at the expense of Q-learning

- RL deficits in the presence of intact RPE signaling
  - Possible with a reduced ability to adaptive modulate learning rate as a function of uncertainty.

- Reduced goal-directed exploration in psychotic illness
  - Can be modeled as a reduced contribution of relative uncertainty to decision making

Models help us think about learning and behavior

- Models are formalized frameworks
  - They can and should be a source of predictions

- There are many possible learning algorithms, because there are many possible learning mechanisms, often operating simultaneously
  - The idea of complementary OFC-driven (fast) and BG-driven (slow) learning system has provided us with many testable hypotheses

- Ideally, parameters correspond to constructs
  - Learning rates
  - Mixing parameters
  - Noise parameters
  - Explore parameters

Caveats

- Important to tailor tasks to severely-mentally-ill population
  - Difficult to model behavior that is unsystematic/random

- Important to characterize subgroups of patients whose data can be fit vs. those whose data *cannot* be fit

- Modeling results should be consistent with behavior

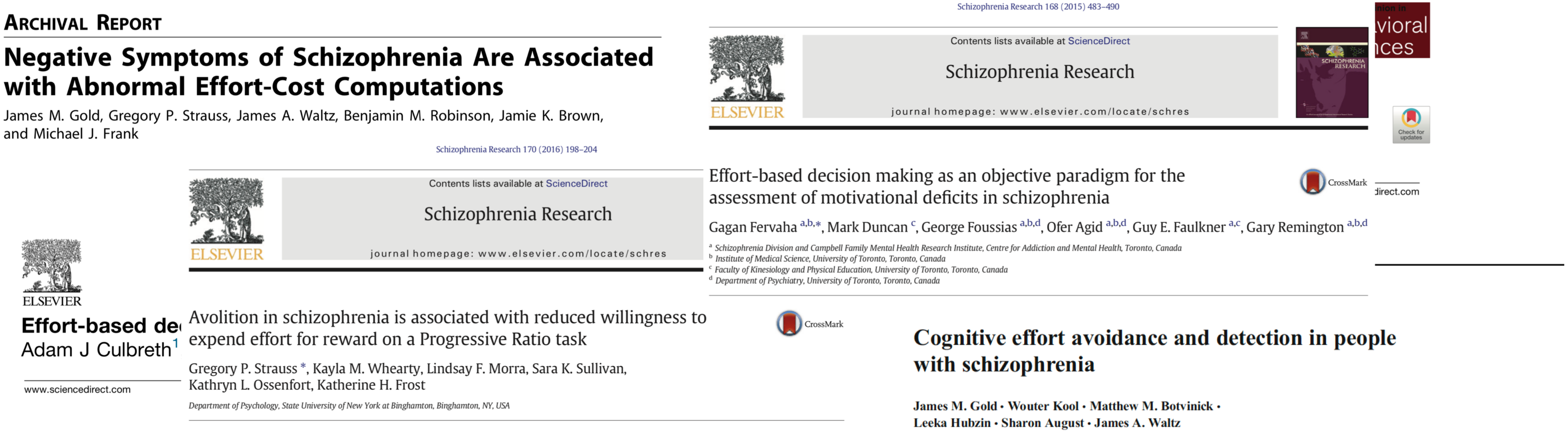
- Important to compare plausible models, based on fit parameters, and demonstrate that the model you report on actually did the best

VII. What do we still want to know?

Many things

How do we account for associations between negative symptom severity and problems of learning about effort cost?

- ➢One path to avolition is if action-value associations are weak
- ➢Another path is if action-cost associations are *strong*



ARCHIVAL REPORT

Negative Symptoms of Schizophrenia Are Associated with Abnormal Effort-Cost Computations

James M. Gold, Gregory P. Strauss, James A. Waltz, Benjamin M. Robinson, Jamie K. Brown, and Michael J. Frank

Schizophrenia Research 170 (2016) 198–204

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

ELSEVIER

Avolition in schizophrenia is associated with reduced willingness to expend effort for reward on a Progressive Ratio task

CrossMark

Gregory P. Strauss *, Kayla M. Whearty, Lindsay F. Morra, Sara K. Sullivan, Kathryn L. Ossenfort, Katherine H. Frost

Department of Psychology, State University of New York at Binghamton, Binghamton, NY, USA

ELSEVIER

Effort-based de

Adam J Culbreth[1]

www.sciencedirect.com

Schizophrenia Research 168 (2015) 483–490

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

ELSEVIER

SCHIZOPHRENIA RESEARCH

Effort-based decision making as an objective paradigm for the assessment of motivational deficits in schizophrenia

CrossMark

Gagan Fervaha [a,b,*], Mark Duncan [c], George Foussias [a,b,d], Ofer Agid [a,b,d], Guy E. Faulkner [a,c], Gary Remington [a,b,d]

[a] *Schizophrenia Division and Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada*
[b] *Institute of Medical Science, University of Toronto, Toronto, Canada*
[c] *Faculty of Kinesiology and Physical Education, University of Toronto, Toronto, Canada*
[d] *Department of Psychiatry, University of Toronto, Toronto, Canada*

vioral
nces

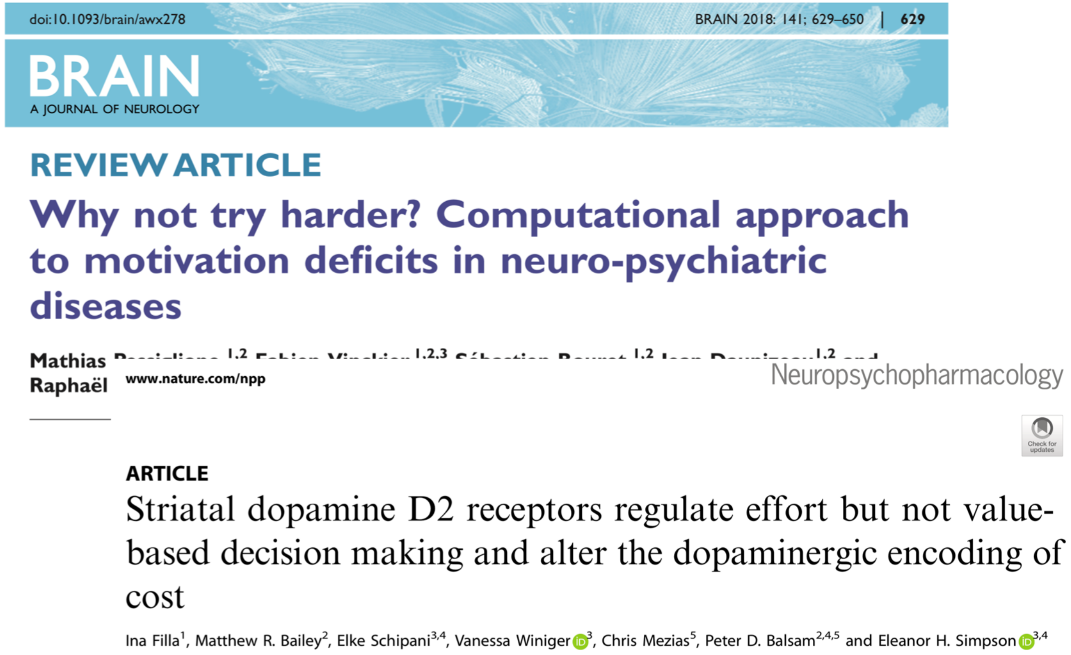
Check for updates

direct.com

Cognitive effort avoidance and detection in people with schizophrenia

James M. Gold · Wouter Kool · Matthew M. Botvinick · Leeka Hubzin · Sharon August · James A. Waltz

How do we account for associations between negative symptom severity and problems of learning about effort cost?



doi:10.1093/brain/awx278 BRAIN 2018: 141; 629–650 | 629

BRAIN
A JOURNAL OF NEUROLOGY

REVIEW ARTICLE

Why not try harder? Computational approach to motivation deficits in neuro-psychiatric diseases

Mathias Pessiglione[1,2] Fabien Vinckier[1,2,3] Sébastien Bouret[1,2] Jean Daunizeau[1,2] and Raphaël

www.nature.com/npp Neuropsychopharmacology

ARTICLE

Striatal dopamine D2 receptors regulate effort but not value-based decision making and alter the dopaminergic encoding of cost

Ina Filla[1], Matthew R. Bailey[2], Elke Schipani[3,4], Vanessa Winiger[3], Chris Mezias[5], Peter D. Balsam[2,4,5] and Eleanor H. Simpson[3,4]

The Journal of Neuroscience, November 19, 2014 • 34(47):15621–15630 • 15621

Systems/Circuits

Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates

The Journal of Neuroscience, June 21, 2017 • 37(25):6087–6097 • 6087

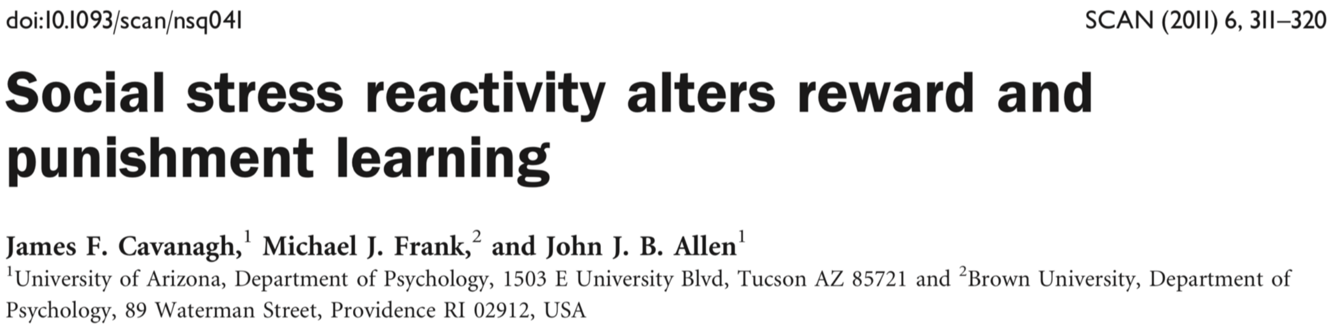
Behavioral/Cognitive

A Selective Role for Dopamine in Learning to Maximize Reward But Not to Minimize Effort: Evidence from Patients with Parkinson's Disease

Vasilisa Skvortsova,[1,2,3] Bertrand Degos,[4] Marie-Laure Welter,[2,3,4] Marie Vidailhet,[4] and Mathias Pessiglione[1,2,3]
[1]Motivation, Brain and Behavior Laboratory, Brain and Spine Institute, Paris, 75013, France, [2]INSERM U1127, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7225, Paris, 75013, France, [3]Université Pierre et Marie Curie, Paris 6, 75013, Paris, France, and [4]Neurology Department, Centre Inter-Régional de Coordination de la Maladie de Parkinson, Hôpital de la Pitié-Salpêtrière, Assistance Publique Hôpitaux de Paris, 75013, Paris, France

- One can estimate prediction errors with regard to cost, just like one can estimate prediction errors with regard to reward.
- The questions of which neural systems underlie the ability to estimate effort cost and the willingness to expend effort are not at all settled

What do we still want to know?
Effects of Stress on Reward- vs. Punishment-driven RL



doi:10.1093/scan/nsq041 SCAN (2011) 6, 311–320

Social stress reactivity alters reward and punishment learning

James F. Cavanagh,[1] Michael J. Frank,[2] and John J. B. Allen[1]
[1]University of Arizona, Department of Psychology, 1503 E University Blvd, Tucson AZ 85721 and [2]Brown University, Department of Psychology, 89 Waterman Street, Providence RI 02912, USA

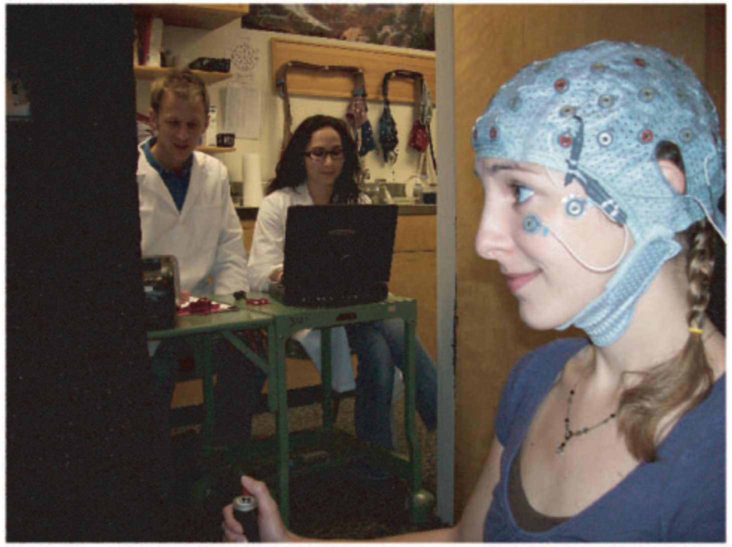


Fig. 2 Depiction of the social evaluative threat stress manipulation during the second performance of the task (T2).

- ➢ Evidence that acute stress shifts balance of Go- and NoGo-learning
- ➢ Stress modulates both dopaminergic and serotonergic pathways, as well as numerous other circuits known to be involved in the processing of rewards and punishments and other salient outcomes

What are the mechanisms by which psychotropic medications influence RL?

- Do antipsychotics decrease learning rates for positive RPEs?
- Do stimulants do the opposite?
- What do SSRIs do?

Acknowledgements

UNIVERSITY *of* MARYLAND
SCHOOL OF MEDICINE

- Jim Gold
- Bob Buchanan
- Zuzana Kasanova
- Jaime Brown
- Rebecca Ruiz
- Adriana Halaby
- Ben Robinson
- Sharon August
- Rebecca Wilbur
- Leeka Hubzin
- Jackie Kiwanuka
- Kim Warren
- MPRC clinical staff

Maastricht University

- Dennis Hernaus

BROWN

- Michael Frank
- Matt Nassar

Berkeley
UNIVERSITY OF CALIFORNIA

- Anne Collins

Duke
UNIVERSITY

- Ziye Xu

Curtin University

- Matt Albrecht

University of Potsdam

- Elliot Brown

NIH National Institute on Drug Abuse

- Elliot Stein
- Tom Ross
- Pradeep Kurup
- Betty Jo Salmeron
- Eliscia Smith
- Angela McNeal
- Carolina Smith
- NIDA clinical staff

Grants

- K12 RR023250-01
- R24 MH72647-01A1
- P30 MH068580-01
- NIDA Residential Research Support Services Contract N01DA-5-9909
- R01 MH080066
- R01 MH094460

Thank you for listening!

jwaltz@som.umaryland.edu