Title

Altered Statistical Learning and Decision-Making in Methamphetamine Dependence: Evidence from a Two-Armed Bandit Task

Permalink

https://escholarship.org/uc/item/48t881zx

Journal

Frontiers in Psychology, 6(DEC)

ISSN

1664-1078

Authors

Harlé, Katia M
Zhang, Shunan
Schiff, Max
et al.

Publication Date

2015

DOI

10.3389/fpsyg.2015.01910

Peer reviewed

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 December 2015
doi: 10.3389/fpsyg.2015.01910

Altered Statistical Learning and Decision-Making in Methamphetamine Dependence: Evidence from a Two-Armed Bandit Task

Katia M. Harlé[1]*, *Shunan Zhang*[2], *Max Schiff*[3], *Scott Mackey*[4], *Martin P. Paulus*[1,5]† and *Angela J. Yu*[2]†

[1] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA, [2] Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA, [3] Department of Psychiatry, Vanderbilt University, Nashville, TN, USA, [4] Department of Psychiatry, University of Vermont, Burlington, VT, USA, [5] Laureate Institute for Brain Research, Tulsa, OK, USA

OPEN ACCESS

Edited by:
Michael Dougherty,
University of Maryland at College Park, USA

Reviewed by:
Christoph T. Weidemann,
Swansea University, UK
Thomas Hills,
University of Basel, Switzerland

***Correspondence:**
Katia M. Harlé
kharle@ucsd.edu

†These authors are co-senior authors on this manuscript.

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: 01 August 2015
Accepted: 26 November 2015
Published: 18 December 2015

Citation:
Harlé KM, Zhang S, Schiff M, Mackey S, Paulus MP and Yu AJ (2015) Altered Statistical Learning and Decision-Making in Methamphetamine Dependence: Evidence from a Two-Armed Bandit Task. Front. Psychol. 6:1910. doi: 10.3389/fpsyg.2015.01910

Understanding how humans weigh long-term and short-term goals is important for both basic cognitive science and clinical neuroscience, as substance users need to balance the appeal of an immediate high vs. the long-term goal of sobriety. We use a computational model to identify learning and decision-making abnormalities in methamphetamine-dependent individuals (MDI, $n = 16$) vs. healthy control subjects (HCS, $n = 16$), in a two-armed bandit task. In this task, subjects repeatedly choose between two arms with fixed but unknown reward rates. Each choice not only yields potential immediate reward but also information useful for long-term reward accumulation, thus pitting exploration against exploitation. We formalize the task as comprising *a learning* component, the updating of estimated reward rates based on ongoing observations, and a decision-making component, the choice among options based on current beliefs and uncertainties about reward rates. We model the learning component as iterative Bayesian inference (the Dynamic Belief Model), and the decision component using five competing decision policies: Win-stay/Lose-shift (WSLS), ε-Greedy, τ-Switch, Softmax, Knowledge Gradient. HCS and MDI significantly differ in how they learn about reward rates and use them to make decisions. HCS learn from past observations but weigh recent data more, and their decision policy is best fit as Softmax. MDI are more likely to follow the simple learning-independent policy of WSLS, and among MDI best fit by Softmax, they have more pessimistic prior beliefs about reward rates and are less likely to choose the option estimated to be most rewarding. Neurally, MDI's tendency to avoid the most rewarding option is associated with a lower gray matter volume of the thalamic dorsal lateral nucleus. More broadly, our work illustrates the ability of our computational framework to help reveal subtle learning and decision-making abnormalities in substance use.

Keywords: Bayesian model, decision-making, reward processing, methamphetamine stimulant, addiction, multi-armed bandit task

INTRODUCTION

Negotiating the tension between exploration and exploitation is an important aspect of cognitive processing, as actions leading to immediate reward may well-conflict with actions that aid the attainment of long-term goals. Daily examples include partying vs. studying, snacking vs. dieting, and relaxing vs. exercising. Understanding how the brain solves this exploration vs. exploitation problem is not only important for basic cognitive science, but also clinical neuroscience, where substance abusers are often faced with the choice between immediate high vs. long-term sobriety. A classical behavioral paradigm used to study the tradeoff between exploration and exploitation task is the multi-armed bandit task (Daw et al., 2006; Behrens et al., 2007; Erev et al., 2008; Gonzalez and Dutt, 2011; Hills and Hertwig, 2012; Zhang et al., 2014), in which subjects must make repeated choices among options (bandit arms) that yield rewards with fixed but unknown probabilities. Selecting different options may either maximize the immediate likelihood of receiving a reward, or the gain in information useful for long-term reward accumulation, thus creating a tension between exploitation and exploration. Bandit problems are widely used in psychology (Steyvers et al., 2009; Zhang and Yu, 2013a), decision neuroscience (Behrens et al., 2007; Cohen et al., 2007), and artificial intelligence (Kaelbling et al., 1996), to study the exploration-exploitation tradeoff. In this work, we use a Bayesian modeling framework to investigate how methamphetamine-dependent individuals differ from healthy controls in performing a two-armed bandit task.

Methamphetamine Dependence and Cognitive Deficits

Methamphetamine dependence (MD) is a serious public health concern (Panenka et al., 2013) associated with a high likelihood of relapse (Brecht and Herbeck, 2014). By 2008, nearly 25 million people worldwide were estimated to have used amphetamine/methamphetamine within the past year (Buxton and Dove, 2008), with abuse being particularly prevalent among younger age groups (Leland and Paulus, 2005). Importantly, executive deficits, most prominent in cognitive control and decision-making paradigms, have been consistently observed in stimulant abusers and implicated in the progression of abuse to dependence (Paulus et al., 2005; Clark et al., 2012; Gowin et al., 2014). Identifying precise neurocognitive markers of such alterations may therefore not only improve our understanding of how neurochemical changes in MD affect decision-making, but it may help identify robust neural predictors of relapse and treatment response.

Reward Processing Impairments in MDI

Although much attention has been given to understanding alterations in impulse control among addicted individuals, neuroimaging studies in both animals and humans point to equally important disturbances in incentive salience and valuation (Goldstein and Volkow, 2011). Specifically, while stimulant abusers exhibit enhanced sensitivity to drugs and drug cues, they show decreased responsiveness to other types of rewards, including secondary reinforcers such as money, which is associated with decreased activation in the orbitofrontal and ventromedial prefrontal cortex (Goldstein et al., 2007; Goldstein and Volkow, 2011). This decreased responsiveness to non-drug rewards is likely to underlie the anhedonic symptoms consistently observed in drug dependence (Koob and Le Moal, 2001), including stimulant dependence (Leventhal et al., 2010), and may promote difficulties regulating stress and negative affect in addicts (London et al., 2004; Tabibnia et al., 2011).

Learning Deficits in MDI

Stimulant dependent individuals also demonstrate impairments in learning new information and in using this knowledge to guide decisions. For instance, chronic amphetamine abusers are not as efficient at learning to avoid high penalty options in the Iowa Gambling Task (Rogers et al., 1999; van der Plas et al., 2009), a deficit shown to be proportional to years of abuse (Rogers et al., 1999). Consistent with this poor learning and difficulties "seeing the big picture," MDI demonstrate a greater discounting of delayed rewards (Hoffman et al., 2006; Monterosso et al., 2007) and a more "myopic" strategy in prediction tasks, with stronger reliance on previous trial outcomes relative to the overall success rates of choice alternatives (Paulus et al., 2002, 2003). Interestingly, during risky decision-making, MDI also demonstrate hypo-activations in the dorsolateral prefrontal cortex (DLPFC) and anterior insula (Ersche et al., 2005; Paulus et al., 2005), brain regions playing an important role in supporting learning and retrieval of stimulus-response associations (Miller and Cohen, 2001; Bunge et al., 2005) and interoceptive function (Paulus and Stein, 2006), respectively.

Together, these findings suggest that MD is associated with impaired tracking and updating of action values and changing contingencies in the environment, which may promote more rigid decision-making strategies (Aron and Paulus, 2007). Combined with reward processing alterations, MDI might also be more likely to make suboptimal choices in complex multi-option environments. Disentangling and quantifying the respective impact of such deficits in these different components of choice behavior (e.g., learning vs. decision policy vs. reward salience) remains a challenge, given only coarse behavioral performance measures (such as monetary earnings or average choice probabilities). In contrast, a model-based approach with more sophisticated representation of individuals' internal computations and variables can perhaps uncover more subtle effects.

A Bayesian Approach to Understanding Decision-Making Deficits in MDI

We have proposed that two separable computational components underlie human choice behavior in the bandit task (Zhang and Yu, 2013a,b): a learning component, the updating of internal knowledge and uncertainty based on successive observations (e.g., successes or failures to obtain reward from the chosen options); and a decision-making component, the selection of an action based on current beliefs and uncertainties about the reward availability at different options). Consistent with this framework, we hypothesize that alterations in one or both of these components could be observed in individuals with a substance disorder such as MDI.

Bayesian models provide a way to address the learning component by quantifying individuals' beliefs about their environment and the associated uncertainty. In this framework, decision-makers are assumed to continuously update their beliefs of the environment based on each new observation. Specifically, we can model statistical learning about reward rates using a version of the Dynamic Belief Model, or DBM (Yu and Cohen, 2009), which assumes subjects believe that environmental statistics can undergo discrete, unsignaled changes without warning. Although reward rates are actually fixed (but unknown) in the bandit task employed in this study, we surmise that individuals may still exhibit *sequential effects, a* persistent tendency to form expectations about upcoming stimuli based on recent trials, which we have shown to arise from the belief that environmental statistics are changeable rather than fixed (Yu and Cohen, 2009). We have shown that DBM accurately predicts sequential effects in a wide variety of behavioral tasks in which stimulus statistics are fixed: perceptual decision-making (Yu and Cohen, 2009), visual search (Yu and Huang, 2014), inhibitory control (Ide et al., 2013; Harlé et al., 2014), and multi-arm bandit tasks (Zhang and Yu, 2013a,b). To model the decision-making component, this Bayesian formulation can be naturally combined with various decision-making (i.e., action selection) models to infer individuals' learning and decision parameters based on their behavioral data.

The Present Study

In this work, we present data from MDI and healthy comparison subjects (HCS) performing a binary-choice version of the multi-arm bandit task (Robbins, 1952). Each arm has a fixed and initially unknown reward rate (probability of reward per trial), though observers may have prior beliefs about the reward rate, and each observed outcome informs the decision-maker about the reward rate. To quantify the learning and decision-making processes in healthy human subjects and MDI, and to examine any subtle differences in the neural circuitry underlying these processes, we use a Bayesian modeling framework (DBM), in combination with five decision policies previously suggested in the literature (Win-Stay/Lose-Shift (WSLS), ε-Greedy, τ-Switch, Softmax, and Knowledge Gradient).

Given evidence of impaired learning in MDI (Miller and Cohen, 2001; Paulus et al., 2002), we hypothesized that relative to healthy comparison subjects (HCS), MDI would be more reliant on a simple learning-independent strategies (e.g., WSLS) and less reliant on more complex learning-dependent, principled strategies (e.g., Softmax). Moreover, given the reduced reward responsiveness of MDI to non-drug reinforcers (Goldstein and Volkow, 2011), we hypothesize that MDI subjects might have altered reward representation before and/or after observing reward outcomes on chosen options.

MATERIALS AND METHODS

Participants

The UCSD Human Research Protections Program and/or the Veterans Affairs San Diego Healthcare System (VASDHS) Internal Review Board approved the study protocol. All subjects gave written informed consent. Sixteen (40% female; mean age = 35.4) sober MDI were recruited from a 28-day inpatient Alcohol and Drug Treatment Program at the Veterans Affairs San Diego Healthcare System and Scripps Green Hospital (La Jolla, CA). To maintain sobriety during the program, participants were screened for the presence of drugs via urine toxicology. In addition, 16 healthy comparison subjects (HCS; 33% female; mean age = 37.1) were recruited via flyers, internet ads (e.g., Craigslist), and local university newspapers. HCS were selected to be matched in age and IQ with MDI. All subjects completed a clinical interview session behavioral session during which they completed the Bandit Task (these study procedures took place between the third and fourth week of treatment for MDI).

Lifetime DSM-IV Axis I diagnoses (including substance dependence) and Axis II antisocial personality disorder were assessed by experienced interviewers using the Semi Structured Assessment for the Genetics of Alcoholism (SSAGA) (Hesselbrock et al., 1999), a semi-structured interview that allows for quantification of lifetime drug use. Diagnoses were based on consensus meetings with a clinician specialized in substance use disorders (MPP) and trained study personnel. The following were exclusion criteria for all groups: (1) antisocial personality disorder; (2) current (past 6 months) Axis I panic disorder, social phobia, post-traumatic stress disorder, major depressive disorder; (3) lifetime bipolar disorder, schizophrenia, and obsessive compulsive disorder; (4) current severe medical disorders requiring inpatient treatment or frequent medical visits; (5) use of medications that affect the hemodynamic response within the past 30 days such as antihypertensives, insulin, and thyroid medication; (6) current positive urine toxicology test; and (7) history of head injuries with loss of consciousness for longer than 5 min. During evaluation, participants performed the North American Adult Reading Test (NAART; Uttl, 2002) as a measure of verbal intelligence (VIQ).

Bandit Task

Participants completed 20 bandit games of 16 trials each on a computer. For each game, participants had 16 tokens (stacked in the middle of the screen) and had to assign one token on each trial to one of the two lottery arms. After placing each token, they either earned one point if the token turned green or zero points if the token turned red see **Figure 1A**. The reward rate for each arm was independent and identically sampled from a Beta distribution ($\alpha = \beta = 2$) at the beginning of each game. In practice, the two arms always had different reward rates in each game, even though on average they had the same mean reward rate (0.5) and standard deviation (0.22). Participants were instructed at the beginning of the experiment that the rewards probability for the arms were independent, and were redrawn at the beginning of each game. They were further instructed to try to maximize the points earned over all trials and in all games. To additionally motivate the participants, we compensated subjects with a dollar amount proportional to their total points earned across all games at the end of the experiment (amounts paid ranged from 6 to $11).

Modeling

We modeled trial-by-trial learning in humans using a form of the hidden Markov model, which we call the Dynamic Belief Model (DBM), which assumes the environmental statistics (i.e., reward rate for an arm in this task) to undergo unsignaled changes (Yu and Cohen, 2009; Zhang and Yu, 2013a,b). It includes the stationary case (the true experimental design) as a special case, whereby the probability of reward rate changing on each trial is exactly 0. We modeled the decision component using five competing decision policies: WSLS, ε-Greedy, τ-Switch, Softmax, and Knowledge Gradient. In the following, we first describe the statistical learning model, then the decision policies, and finally the model-selection procedure for identifying individual decision strategies and group learning parameters.

Dynamic Belief Model

As a simple variant of a hidden Markov model, the generative model assumes that on each game, the two arms have reward rates, θ_{m}, $m = 1$ or 2, each independently generated from the generic Beta prior distribution $q^0(\theta_{\mathrm{m}}) = \text{Beta}\,(\alpha_0, \beta_0)$ with mean $r = (\alpha_0)/(\alpha_0 + \beta_0)$. For simplicity, we assumed that the sum of its two parameters (which controls variance) to be fixed at $\alpha_0 + \beta_0 = 4$. We also assumed that the reward rate for each arm has a probability γ of staying the same as last trial, and a probability (1-γ) of being independently reset and re-drawn from q^0 on any trial, hence embodying the assumption of non-stationarity in the Dynamic Belief Model see **Figure 1B**. We call γ the stability parameter, since larger γ results in a more stable arm that changes reward rates less frequently ($\gamma = 1$ is a special-case arm that never changes reward rate at all).

Given the above generative model, we can use standard Bayesian probability theory to compute the posterior distribution over reward rates of the two arms on each trial, after making an observation. We use the notation $q_m^t\,(\theta_m^t) := \Pr(\theta_m^t|\mathrm{x}^t)$ to denote the posterior probability distribution over the reward rate for the mth arm on the tth trial, denoted θ_m^t, given the observed sequence of successes and failures from all previous trials, denoted $\mathrm{x}^t := (\mathrm{x}^1 \ldots \mathrm{x}^t)$. On each trial t, the observer's iterative prior distribution marginalizes over uncertainty about whether there has been a reward rate change on the current trial, and is therefore a mixture of last trial's posterior and the generic prior:

$$\Pr(\theta_m^t = \theta|\mathrm{x}^{t-1}) = \gamma q_m^{t-1}(\theta) + (1-\gamma)q^0(\theta)$$

To update the posterior after the current trial, for the chosen arm only (assuming it is the mth arm), having observed the outcome R_m^t (1 for a reward, 0 for no reward), the new posterior distribution for the chosen arm can be computed via Bayes' rule:

$$q_m^t(\theta_m^t) \sim \Pr(R_m^t|\theta_m^t)\Pr(\theta_m^t|\mathrm{x}^{t-1})$$

whereas, the posterior for the un-chosen arm is the same as the prior at the beginning of the current trial (since there has been no new observation). The mean of the prior distribution, μ_m^t, is what we call *estimated reward rate* for arm m.

In the actual experimental design, the reward rates were fixed. This is one possible, special case setting also captured by the DBM, by assuming the probability of the reward rate changing on any trial is 0 ($\gamma = 1$), which we call the Fixed Belief Model (Yu and Cohen, 2009; Zhang and Yu, 2013b).

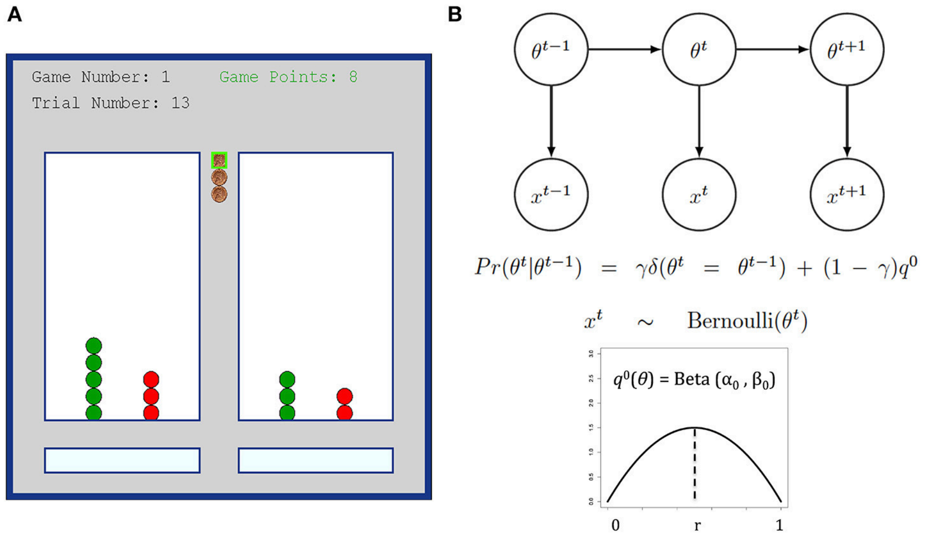

A
Game Number: 1
Game Points: 8
Trial Number: 13
B
θ^{t-1}
θ^t
θ^{t+1}
x^{t-1}
x^t
x^{t+1}
$Pr(\theta^t|\theta^{t-1}) = \gamma\delta(\theta^t = \theta^{t-1}) + (1-\gamma)q^0$
$x^t \sim \text{Bernoulli}(\theta^t)$
$q^0(\theta) = \text{Beta}\,(\alpha_0, \beta_0)$
0
r
1

FIGURE 1 | (A) Bandit task interface snapshot. Participants completed 20 games, each with 16 trials. **(B)** DBM illustration and the generative equations. The reward rate of each arm are assumed to be independently drawn at the start of a game from a Beta distribution $q_0 = \text{Beta}\,(\alpha_0, \beta_0)$, fixed throughout the game, and with mean $r = (\alpha_0)/(\alpha_0 + \beta_0)$. DBM assumes that subjects believe that the reward rate θ for any arm can reset on any trial with probability 1-γ, otherwise it is the same value as the last trial.

Decision Policies

In the cognitive science and reinforcement learning literatures, a number of decision policies with varying levels of complexity have been used to model human bandit choice (Daw et al., 2006; Steyvers et al., 2009; Zhang et al., 2014). These policies can be conceptualized as underlying the choice of a goal-directed action based on the individual's current beliefs and knowledge of their environment. Here, we considered five models, ordered below by increasing complexity: Win-stay/Lose-shift (WSLS), τ–Switch, ε-Greedy, Softmax, and Knowledge Gradient. WSLS is a simple, learning-independent heuristic policy, which stays with the last chosen arm after a reward with probability γ^{w} and switches to the other arm after a loss (no reward) with probability γ^{l} (Robbins, 1952). τ-*Switch* is another learning-independent policy that assumes that the decision-maker uses a fast-and-frugal heuristic for their choice selection depending on the counts of previous successes and failures for both arms. They choose randomly when the two arms have equal counts of previous successes and failures, namely $S_1 = S_2$ and $F_1 = F_2$. The second situation is when one arm is better/worse than the other; for example, Arm 1 is better (or Arm 2 is worse) if $S_1 > S_2$ while $F_1 \leq F_2$, or $F_1 < F_2$ while $S_1 \geq S_2$, and the model chooses the better arm with (a large) probability γ^{τ}. When one arm has both more previous successes and failures than the other, it is an "exploit" option, and the other arm is an "explore" option; in this situation, the model chooses the explore option with probability γ^{τ} if the current trial is before the switch point τ, otherwise chooses the exploit option with probability γ^{τ} if the current trial is after τ. A detailed description of this model can be found in Lee et al. (2011). γ^{τ} is a parameter of "accuracy of execution," which captures the proportion of choices that are consistent with either the exploration or exploitation policy on any given trial (Steyvers et al., 2009). ε-Greedy assumes that one chooses the alternative with the greatest estimated reward rate with probability 1-ε on each trial, but chooses randomly among the remaining arms with probability ε (Barto, 1998). The Softmax decision policy assumes that the decision-maker chooses among the options with probabilities related to the inferred reward rates of the respective arms, but often with exaggerated ratios (over-matching) compared to the estimated reward rates, but typically not linearly (Luce, 1959). Here, we assumed that the choice probabilities are normalized polynomial functions of the estimated reward rates, with polynomial parameter *b*, e.g., Pr(choosing arm 1) = $\mu_1^{b}/(\mu_1^{b} + \mu_2^{b})$, so that when *b* approaches infinity, the maximally rewarding option is always chosen (maximizing), when *b* is 1, it is probability matching, and when *b* is 0, the arms are chosen randomly (with equal probability. Knowledge Gradient or KG (Frazier et al., 2008; Ryzhov et al., 2012; Zhang and Yu, 2013b) is the most sophisticated among the heuristic policies we consider here. In its original formulation, KG is a deterministic policy that chooses the arm with the highest combined gain of immediate reward (first term in the equation below) and longer-term "knowledge gain" (second term), with the linear tradeoff parameterized by the distance to horizon (e.g., fewer trials left results in less emphasis on "knowledge gain"). Namely, the decision rule is

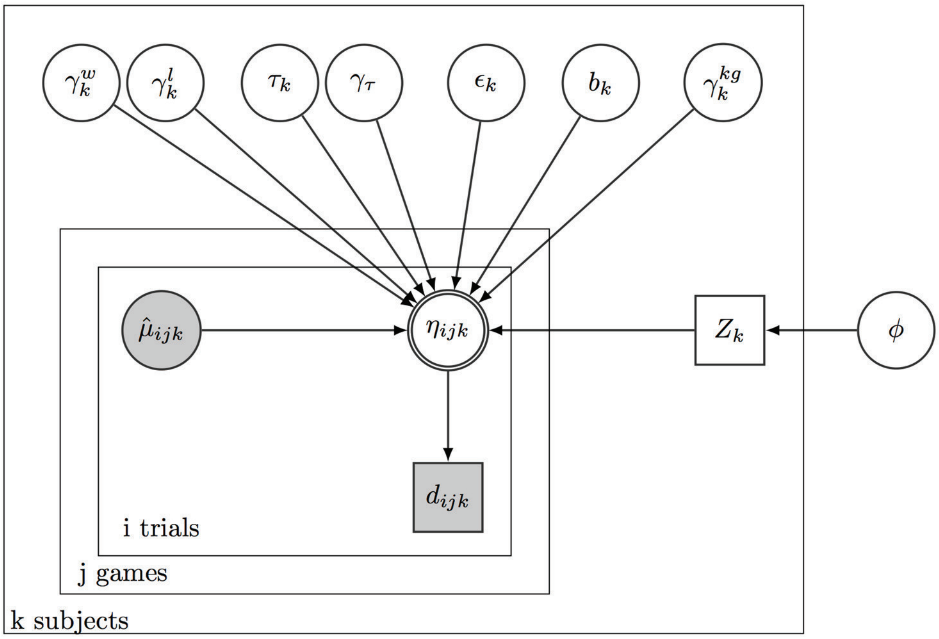
$$D^{KG,t} = argmax_m \mu_m^t + (T - t - 1) v_m^{KG,t}$$

where $v_m^{KG,t} = \mathrm{E}[max\, \theta^{t+1} | D^t = m, q^t]\text{-}max\, \theta^t$ is the approximate value function for choosing arm *m* on trial *t*, under the current belief state q^t. This formulation is similar to the optimal policy except that the second term in *KG* approximates the value of exploration (second time) by using a lower-bound, that attained by allowing only one more exploratory step and exploitation thereafter, whereas the real optimal policy would also consider the possibility of further exploratory choices, which involves a much more expensive computation. Note that *KG* does not actually choose to exploit after one more exploratory step, it merely estimates the *value* of further exploration using this computational assumption. It is interesting to note that *KG* is equivalent to the optimal policy of explicitly maximizing cumulative gain, when the reward rates are assumed to be fixed ($\gamma = 1$) and there are only two arms (Frazier et al., 2008), however, they are not equivalent in the problem here under the DBM (non-stationary) assumptions. In this work, we extended the original formulation of *KG* by adding a free parameter γ^{kg} that turns this deterministic policy into a probabilistic policy, such that Pr(choosing arm $m | D^{KG,t} = m) = \gamma^{kg}$, so as to match the other algorithms better in terms of the number of free parameters. Both *KG* and the optimal policy increasingly favor exploitation over exploration with fewer trials left; as we did not see this tendency in subjects' behavior in both a previous study (Zhang and Yu, 2013b), and the current data set (results not shown), we do not explicitly consider the optimal policy here. Detailed description of the optimal policy that KG approximates under stationary bandit setting can be found in Zhang and Yu (2013b).

Bayesian Model Comparison

As a compromise between model specificity and statistical power, we assumed individuals in the same group (i.e., MDI or HCS) to share the same DBM stability parameter γ and the mean of the generic prior *r*. However, we assumed that individuals may differ in their decision policy, both in terms of which policy and what parameter setting. Specifically, for each fixed pair of DBM parameter values (γ, *r*), where γ and *r* each varies from 0 to 1 in increments of 0.1 (they are bounded by 0 and 1), we can compute for every subject a sequence of trial-wise prior distributions over reward rates for the two arms, based on his/her actual sequence of choices and observations (reward or no reward), and thus a likelihood of observing the subject's choice on each trial for each policy (and each parameter setting of each policy). This would allow us to compute a joint likelihood of all data (all choices of all subjects) by multiplying the likelihood of each observation (a subject's choice on one trial), and thus a means for comparing estimates of (γ, *r*) between MDI and HCS, and the decision policies and policy parameters across subjects.

However, it is very computation- and memory-intensive to compute a joint likelihood for all model parameters (e.g., by



$$\eta_{ijk} = \begin{cases} \gamma_k^w & \text{if previous success on arm 1 and } Z_k = 1 \\ 1 - \gamma_k^w & \text{if previous success on arm 2 and } Z_k = 1 \\ \gamma_k^l & \text{if previous failure on arm 1 and } Z_k = 1 \\ 1 - \gamma_k^l & \text{if previous failure on arm 2 and } Z_k = 1 \\ 1/2 & \text{if two arms are equal and } Z_k = 2 \\ \gamma_\tau & \text{if arm 1 is ``better'' and } Z_k = 2 \\ 1 - \gamma_\tau & \text{if arm 1 is ``worse'' and } Z_k = 2 \\ \gamma_\tau & \text{if arm 1 is explore-choice and } i < \tau_k \text{ and } Z_k = 2 \\ 1 - \gamma_\tau & \text{if arm 1 is explore-choice and } i > \tau_k \text{ and } Z_k = 2 \\ \gamma_\tau & \text{if arm 1 is exploit-choice and } i > \tau_k \text{ and } Z_k = 2 \\ 1 - \gamma_\tau & \text{if arm 1 is exploit-choice and } i < \tau_k \text{ and } Z_k = 2 \\ 1/2 & \text{if } \hat{\mu}_{ijk}^{(1)} = \hat{\mu}_{ijk}^{(2)} \text{ and } Z_k = 3 \\ 1 - \epsilon_k & \text{if } \hat{\mu}_{ijk}^{(1)} > \hat{\mu}_{ijk}^{(2)} \text{ and } Z_k = 3 \\ \epsilon_k & \text{if } \hat{\mu}_{ijk}^{(1)} < \hat{\mu}_{ijk}^{(2)} \text{ and } Z_k = 3 \\ (\hat{\mu}_{ijk}^{(1)})^{b_k} / ((\hat{\mu}_{ijk}^{(1)})^{b_k} + (\hat{\mu}_{ijk}^{(2)})^{b_k}) & \text{if } Z_k = 4 \\ \gamma_k^{kg} & \text{if original KG decision rule chooses arm 1 and } Z_k = 5 \\ 1 - \gamma_k^{kg} & \text{if original KG decision rule chooses arm 2 and } Z_k = 5 \end{cases}$$

$$Z_k \sim \text{Categorical}(\phi)$$
$$d_{ijk} \sim \text{Bernoulli}(\eta_{ijk})$$

FIGURE 2 | Bayesian model comparison. We simultaneously infer the latent model usage and model parameters for all five different strategic decision-making models, including two learning-independent heuristic models (WSLS, τ-Switch), two learning-dependent heuristic models (ε-Greedy, Softmax), and Knowledge Gradient, based on observed data (subjects' actual choices and outcomes). See Materials and Method for more details. In the Bayesian graphical model, a node is double-bordered if it is deterministic (on its parents), otherwise stochastic; a node is gray if it is known/observed, otherwise unknown and to be inferred. Circle node is continuous, whereas rectangular node is discrete. η is the probability for choosing arm 1 for subject k, at trial i of game j. This choice probability is deterministically dependent on the estimated reward rates and the decision policy.

discretizing each parameter) for all data. We therefore specified a Bayesian graphical model of how the data (subjects' observed choices) are generated from the underlying model parameters see **Figure 2**, and use WinBUGS (Spiegelhalter et al., 2003), to sample (using Markov chain Monte Carlo or MCMC) from the full joint posterior distribution over all the decision-policy parameters conditioned on the observed data. Specifically, we assumed each subject k utilizes policy l, $z_k = l$, for all trials with a categorical probability distribution that has a Dirichlet prior distribution, Dir(1,1,1,1,1), which yields a marginal prior probability of 1/5 for each policy, and each decision policy parameter has a prior distribution that is uniform over the unit interval (γ^w, γ^l, ε, and γ^τ) except for b, the Softmax parameter, which has a very flat Gamma prior with support over R^+ (mean = 10, std = 10). Because all the priors are flat (or nearly flat), the posterior surface is approximately proportional to the data likelihood, but represented by samples, such that the likelihood (posterior probability) for a region of the parameter space is reflected in the relative number of samples. We compared the "average" likelihood of different settings of (γ, r) for each group (by marginalizing over the uncertainty associated with the choice of decision policies and their parameters), or the "average" likelihood of each policy for every subject (by marginalizing over the uncertainty associated with the parameters of each policy). This constitutes a form of Bayesian model comparison, which has the convenient feature of automatic penalization of model complexity.

Based on the specified generative model, for each setting of (γ, r), we obtained from WinBUGS 2000 samples (two MCMC chains, each containing 1000 samples with a burn-in period of 1000 samples, and using standard checks for convergence, Gelman and Rubin, 1992) of model parameters, each sample containing the setting of the indicator variable z_k specifying subject k's decision policy, and the parameter setting of the relevant policy; WinBUGS also returns the data likelihood associated with each sample. To identify the DBM parameters (γ, r) for each group (MDI or HCS), we computed the marginal data likelihood by integrating out uncertainty over z_k and all decision policy parameters [i.e., adding up the likelihood of all samples for (γ, r)] and use the maximal marginal likelihood estimates for (γ, r). Having fixed (γ, r) for each group (MDI or HCS), we could then identify the policy used by each subject, $\hat{z}_k$, by again finding the maximal marginal likelihood estimates for z_k (i.e., adding up the number of samples for each setting of the categorical variable z_k). Finally, to identify the decision policy parameter used by subject k, we considered all samples where $z_k = \hat{z}_k$, and found the policy parameter setting (sample) with the highest likelihood.

Group Comparison Statistics

For behavioral variables with repeated measures (e.g., trial-wise reaction times and game points), we fit hierarchical generalized mixed-effect linear models treating subject as a random factor (with varying intercepts in one model, and also varying slopes in another model) and other variables as fixed effects (Baayen et al., 2008). For group comparison of individually fit parameters, independent two-sample t-tests were used. To compare the learning (DBM) parameters (γ, r) between the two groups, we estimated the "mean" and "variance" of γ, r, by resampling from the marginal data likelihood (shown as grayscale maps in **Figure 3**), or equivalently the marginal posterior distribution assuming a uniform prior over (γ, r), using one MCMC chain of length 10^7. The mean was estimated by the sample mean, the variance was estimated by the sample variance. These estimates were then used, together with the actual group sizes (16) for MDI and HCS, to construct the 95% CI (t-critical value of 2.04 for a two samples t-test; df = 16+16-2 = 30).

MRI Image Acquisition and Voxel-Based Morphometry (VBM)

High-resolution *in vivo* structural MR-images (T_1-weighted spoiled gradient recalled [SPGR] imaging, TR = 8 ms, TE = 3 ms, slices = 172, FOV = 25 cm, approximately 1 mm^3 voxels) were acquired in all subjects on a 3.0 Tesla Signa EXCITE scanner (GE Healthcare, Milwaukee, WI). Optimized voxel-based morphometry (VBM) was performed with the FSL-VBM pipeline (Douaud et al., 2007); http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLVBM) using FSL tools (FSL-4.1.6; (Smith et al., 2004). Optimized VBM uses an iterative approach to segmentation and normalization that results in a more accurate identification of gray and white matter (Good et al., 2001). Brains were first automatically extracted from the skull with BET (Smith, 2002). Tissue-types (i.e., gray matter, white matter, CSF) were then segmented with FAST4 (Zhang et al., 2001). The gray matter images were aligned with the MNI-152 standard space by affine registration (d.f. = 12) using the FLIRT tool (Jenkinson et al., 2002) followed by non-linear registration using FNIRT (Andersson et al., 2007) and averaged to create a study-specific template. Segmented gray matter images in native space were then re-registered to this template. To preserve information about absolute volume, partial volume images were modulated by the non-linear component of the Jacobian determinants generated during spatial normalization thus obviating the need to correct for total intracranial volume (Scorzin et al., 2008). To make the residuals in subsequent analyses conform more closely to a Gaussian distribution and to account for individual differences in brain anatomy, the modulated GM images were smoothed with an isotropic Gaussian kernel, $\sigma = 3\,\text{mm} \approx 7.06\,\text{mm}$ FWHM. The average modulated gray matter volume was extracted from 70 cortical and subcortical regions of interest (ROIs). The ROIs were defined by a maximum probability map based on the Talairach atlas. The construction of these ROIs are described elsewhere (Fonzo et al., 2013; Ball et al., 2014).

RESULTS

Behavioral Measures

Figure 1A illustrates the bandit task we used in this study. Combining all subjects, we found a negative linear relationship between reaction times (RT) and game number ($B = -12$ ms, $t = -3.6$, $p < 0.001$, model omnibus test: $\chi^2 = 35.9$, $p < 0.001$; Mean RT = 1428 ms), such that individuals made faster decisions as they had more experience with the bandit task. However, neither the group main effect ($\chi^2 = 0.46$, $p = 0.50$) nor the

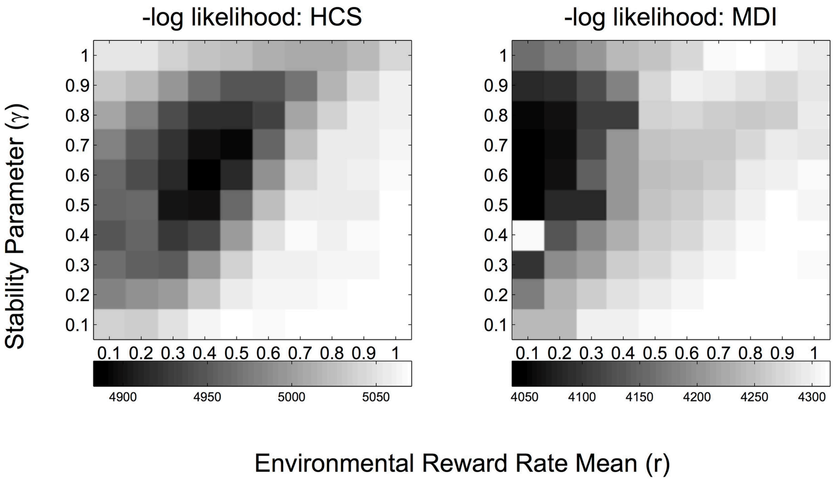


FIGURE 3 | -Log likelihood grayscale plots for each pair of DBM parameter values (γ, *r*) fitted at the group level (MDI, methamphetamine dependent individuals; HCS, healthy comparison subjects). Values represent −2*log likelihood, where the likelihood is marginalized over uncertainty associated with the decision policy utilized by each subject and the parameter setting of the policy. Darker color indicates lower log likelihood, thus better fit.

group × game interaction ($\chi^2 = 0.86, p = 0.35$) were significant, i.e., MDI did not differ from HCS on their general latency to select an option nor on their decrease in latency for later trials. In general, we did not find earnings to vary significantly as a function of game number ($\chi^2 = 0.02$, $p = 0.89$; Mean Game Points = 8.9)—in particular, subjects did not improve in their performance as they had more experience. Correspondingly, we also found no group difference in total earnings, both in terms of overall group effect ($\chi^2 = 0.01$, $p = 0.93$) or group × game interaction ($\chi^2 = 0.03$, $p = 0.86$). Thus, MDI and HCS had similar overall performance in the bandit task.

Learning Model

The best data-fitting parameters for the learning model (DBM, see **Figure 1B**) were inferred for each group: these consist of estimates for the prior expectation of reward rate, $\hat{r} = (\hat{a})/(\hat{a} + \hat{b})$, and the stability parameter, $\hat{\gamma}$, which also controls the effective exponential memory window size and thus can be thought of as a discount rate parameter (larger γ = less assumed volatility in reward rates = longer memory window = slower/less discounting, see Yu and Cohen, 2009). **Figure 3** shows the logarithm of the marginal likelihood values for each setting of (γ, r) where each variable varies between 0 and 1 in increments of 0.1, i.e., how well different settings of (γ, r) can account for all the observed choices of all subjects in a group (MDI or HCS), after marginalizing over uncertainty about the hidden parameters that specify which decision policy each subject uses and the parameter setting of that policy (see Materials and Methods for details). As shown in **Figure 3**, we found that MDI and HCS have similar estimated stability parameter $\hat{\gamma}$ (MDI: Mean = 0.60; HCS: Mean = 0.59; CI95%: $-0.005 < \mu_{\text{HCS}}-\mu_{\text{MDI}} < 0.006$). Both HSC and MDI behave as though they believe the environment to be changeable—in fact, at approximately once every $1/(1\text{-}\hat{\gamma}) = 1/(1\text{-}0.6) = 2.5$ trials—instead of assuming the reward rates to be static, which was the "true" experimental design. The estimated ∩γ is smaller than values found in most other tasks, which tends to be around 0.7–0.8 (Yu and Cohen, 2009; Ide et al., 2013; Yu and Huang, 2014; Zhang et al., 2014; Ma and Yu, 2015), and may potentially be due to the longer inter-trial interval used in this task (temporal discounting may be influenced also by absolute time, not only discrete trials as assumed in DBM).

Unlike the estimates for stability/volatility, we found that MDI and HCS *do differ* in their prior beliefs about mean reward rate (MDI: Mean = 0.11; HCS: Mean = 0.40; CI95%: $0.28 < \mu_{\text{HCS}}-\mu_{\text{MDI}} < 0.31$), such that MDI seem to have lower prior belief of receiving reward from any lottery arm (probability = 0.11) relative to HCS (probability = 0.40), i.e., MDI individuals are overall more pessimistic about the same reward environment than HCS.

Decision Policy

To identify the decision policy utilized by each individual, and to estimate the relevant model parameter(s), we first fixed the DBM parameters at the values estimated for each group (see previous section), found the policy for each individual that yield the highest marginal likelihood for the observed choice data for that subject (marginalized over parameter settings), and then estimated the policy parameter setting that achieves the highest likelihood. We found that WSLS or Softmax to best explain each participant's data, or in other words, Knowledge Gradient, τ-switch, and ε-greedy are not as good at predicting any participant's behavioral data. While WSLS and Softmax were each found to be the best fitting policy for some individuals in each group, there is was a statistical trend toward a higher proportion of MDI relying on the learning-independent WSLS (8/16 = 50%) compared to HCS (only 3/16 = 19%; $\chi^2 = 3.46$, $p = 0.06$). Conversely, while a majority of HCS used a Softmax strategy (81%), only half of MDI used such model (50%; see **Figure 4A**).

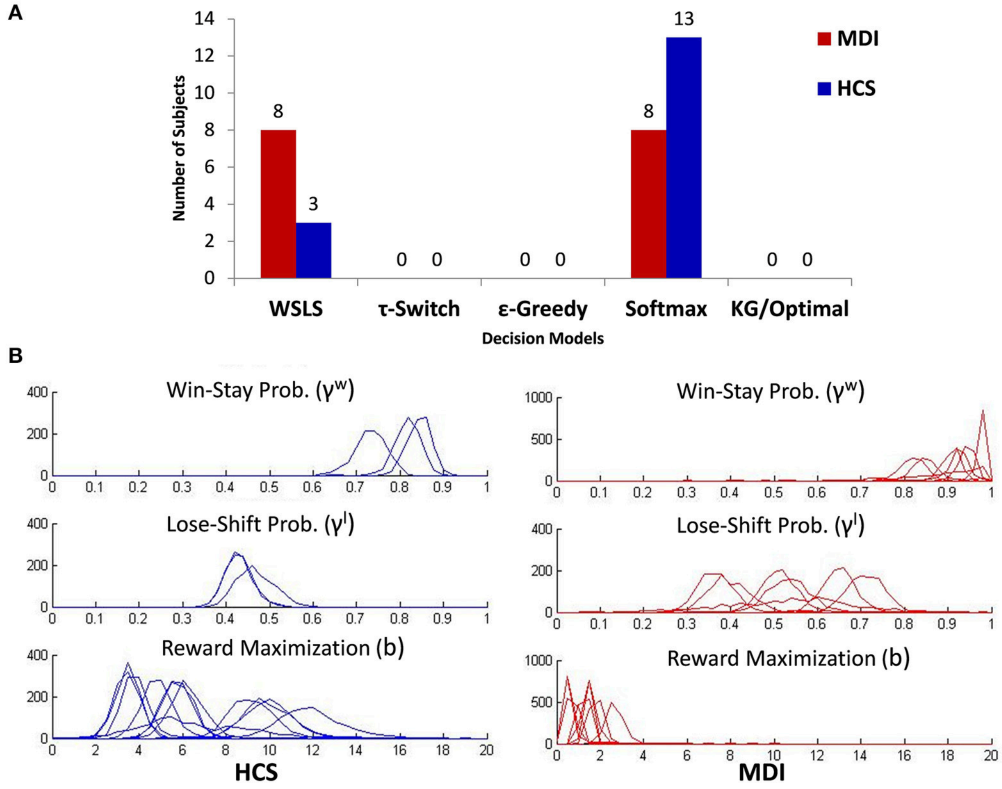


FIGURE 4 | (A) Classification of all subjects by group (MDI, methamphetamine dependent individuals; HCS, healthy comparison subjects) and model usage based on posterior mode of the z_{ijk} parameter. **(B)** Posterior distributions for the parameters associated with the strategic control model used by participants: Win-Stay/Lose-Shift (WSLS; γ^w γ^l: top two graphs) and Softmax (*b* parameter; bottom graph). The y-axis depicts a histogram of the counts of MCMC samples obtained for each of these parameters for each individual (overlaid) identified to be most likely to be using the corresponding policy.

Based on this result, we only provide group comparisons of the parameter values of the two best-fitting models (γ^w and γ^l for WSLS; and b for Softmax). **Figure 4B** shows the posterior distribution over model parameters (assuming uniform priors) for different subjects (overlaid), where each figure only contains the posterior of individuals whose choices are best explained by the corresponding policy. In practice, the posterior distributions are approximated by the counts of MCMC samples obtained by WinBUGS for each of these parameters for each individual (see Materials and Methods). In the following, we used only maximum-likelihood estimates (MLE) for decision policy parameters. Among individuals using WSLS, MDI, and HCS did not differ in their tendency to lose-shift, i.e., switching arms after not receiving a reward (MDI: Mean $\gamma^l = 0.54$; HCS: Mean $\gamma^l = 0.45$, $t_{(9)} = 1.2$, $p = 0.25$), but MDI had significantly higher tendency than HCS to win-stay, i.e., choosing the same arm after receiving a reward (MDI: Mean $\gamma^w = 0.91$; HCS: Mean $\gamma^w = 0.81$, $t_{(9)} = 2.7$, $p = 0.03$). Among those using the Softmax strategy, relative to HCS, MDI had a significantly lower reward maximization parameter (MDI: Mean $b = 1.58$; HCS: Mean $b = 6.99$, $t_{(19)} = 5.6$, $p < 0.001$), indicating that MDI select actions more like *matching*, while HCS act more like *maximizing*, which has also been found for healthy individuals in other learning and decision-making task (Zhang et al., 2014).

Relationship Between Computational Parameters and Gray Matter Brain Volumes

To further investigate the potential neural substrate of individual behavioral differences, we conducted exploratory correlational analyses within each set of model users (i.e., WSLS and Softmax users) between individual parameter values and average VBM gray matter relative volumes for 70 ROIs. No association was found between WSLS parameters (γ^w and γ^l) and anatomical gray matter volumes. Within Softmax users, we found a positive association between the reward maximization b parameter and gray matter volumes of the thalamic lateral dorsal nucleus (LD; $r = 0.45$, $p < 0.05$; see **Figure 5A**).

Given the strong relationship between clinical group status and Softmax *b* parameter, we further investigated the relationship between clinical status and gray matter volume in this thalamic region and the potential mediating role of the Softmax parameter in this relationship. To do so, we used a hierarchical regression method, with both linear and logistic regressions to accommodate for the dichotomous clinical group variable (Baron and Kenny, 1986). As expected, a first model showed that higher

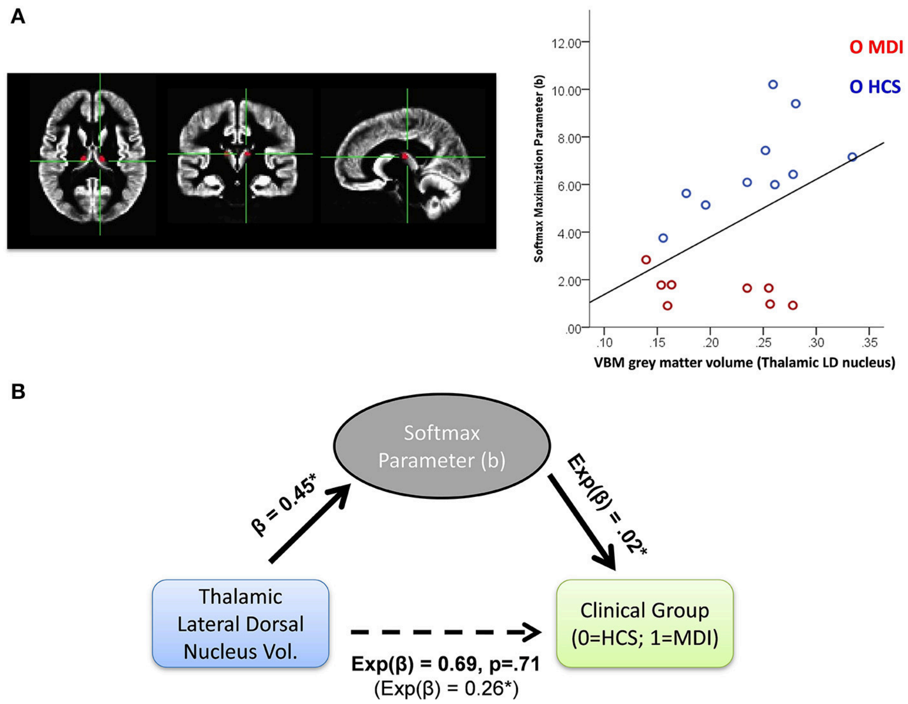


FIGURE 5 | (A) Left panel: thalamic lateral dorsal nucleus (LD) ROI in axial, coronal, and sagittal planes; Right: correlation between Softmax b paramter and VBM based gray matter volume in the thalamic LD in Softmax users ($N = 17$). **(B)** Mediation analysis with hierarchical generalized linear models (with a logit link function for models predicting clinical group). Softmax *b* parameter was found to fully mediate the negative relationship between thalamic LD gray matter volume and likelihood to be MDI vs. HCS (MDI, methamphetamine dependent individuals; HCS, healthy comparison subjects); $^*p < 0.05$.

Softmax *b* parameter was associated with a lower likelihood to belong in the MDI group (ominibus $\chi^2_{(1)} = 17.1$, $p < 0.001$; odd ratio = 0.02). Another model showed that higher thalamic LD gray matter volume was associated with a lower likelihood to belong in the MDI group (ominibus $\chi^2_{(1)} = 4.9$, $p < 0.05$; odd ratio = 0.26). Thalamic LD gray matter volume was positively related to Softmax b parameter, $F_{(1,\ 18)} = 4.2$, $p = 0.05$ (beta = 0.45). Importantly, adding Softmax b parameter as a second predictor of clinical status removed the effect of thalamic LD gray matter volume ($p = 0.71$), leaving the Softmax parameter as the only significant predictor of clinical status (ominibus $\chi^2_{(1)} = 17.2$, $p < 0.001$; odd ratio = 0.03), consistent with a full meditation of the Softmax parameter (see **Figure 5B**).

DISCUSSION

In this study, we applied a probabilistic learning and decision-making model human choice behavior data in a bandit task, in order to investigate cognitive differences in learning and decision-making between recently sober MDI and HCS. To model the representation and updating of individuals' beliefs, we used the Dynamic Belief Model, a Bayesian iterative inference model which assumes the environment to undergo unpredictable and discrete changes (Yu and Cohen, 2009). The decision-making component was modeled with a set of five well-established decision policies from the cognitive science and reinforcement learning literatures, including Win-stay/Lose-shift (WSLS), ε-Greedy, τ-Switch, Knowledge Gradient, and Softmax. To our knowledge, this is the first study using such hierarchical Bayesian approach to assess reward processing in a clinical population such as MDI.

Bandit Choice Behavior

The bandit task has been a popular behavioral paradigm for studying the exploration-exploitation tradeoff, as the task involves a potential conflict between actions that maximize the short-term potential of immediate reward and the long-term gain of information that maximizes total rewards. Our modeling framework decomposes the task into a learning component, which consists of learning about initially unknown reward rates for the two arms in each game, and a decision component, which consists of choosing an arm on each trial based on previous observations and any prior beliefs. We found that the learning algorithm that best describes each of HCS and MDI subjects (in the latter group, only those who show learning) indicates the subjects to be assuming the reward rate statistics to be changing on a relatively fast timescale (about once every 2.5 trials), despite the experimental reward rates to be actually constant in a game, but consistent with healthy human choice behavior in a variety of behavioral tasks. A consequence of this peculiar non-stationarity belief is that a subject's belief (reflected in his/her choice) on the current trial is strongly influenced by the outcome of the most recent trials in the past, and exponentially less by outcomes farther into the past (Yu and Cohen, 2009), producing what is classically known in psychology as *sequential effects*. In terms of the decision policy, we found that every HCS and MDI subject was best described as utilizing a heuristic Softmax policy or the even the simpler, learning-independent WSLS policy. Unlike the more sophisticated Knowledge Gradient policy (and of course the optimal policy, see Zhang and Yu, 2013b), Softmax does not explicitly assess the relative value of future exploratory gain vs. immediate exploitative gain, but rather uses a single fixed parameter (b in this paper) to heuristically "loosen" up the choice policy relative to the estimated reward rates of the given options. Consequently, one important difference between Softmax and Knowledge Gradient (and also the optimal policy) is that the former policy is insensitive with respect to the number of trials left (known as the horizon in reinforcement learning literature), while Knowledge Gradient (and the optimal policy) weigh exploratory gain less relative to exploitative gain as the horizon gets closer and there is less time left to take advantage of any additional information gained. This insensitivity to horizon finding is consistent with what we previously found for healthy human subjects in the bandit task (Zhang and Yu, 2013b). However, in addition to these coarse similarities between MDI and HCS in both learning and decision-making, there are also some subtle but important differences, as we detail below.

Learning Alterations in MDI

Relatively fewer MDI (50%) than controls (81%) used the Softmax decision policy, instead favoring the WSLS policy. Thus, less MDI were likely to use a learning-supported strategy, which uses estimated reward rates for all arms, and instead used a myopic heuristic relying solely on previous trial outcome. This result is consistent with research suggesting that MDI are impaired in learning and updating their knowledge of the environment and generally have difficulties "seeing the big picture." For instance, along with weaker recruitment of neural regions associated with learning such as the DLPFC and anterior insula, impairments in working memory (Chang et al., 2002) and sequential decision-making (Rogers et al., 1999; Paulus et al., 2002) have been noted in this population, along with difficulties detecting trends and integrating new information to predict future outcomes (Aron and Paulus, 2007).The present results suggest such deficits may also manifest in the type of decision policies implicitly chosen to make reward-based decisions.

Non-drug Reward Hyposensitivity in MDI

We also found that among individuals using a Softmax strategy, MDI had on average lower reward maximization (b) parameter values compared to HCS. Because this parameter reflects the weight given in the action policy to the options with highest predicted reward rates, this individual metric may be conceptualized as an individual's reward sensitivity bias, independently of their expectations of reward. Interestingly, such parameter value tended toward a value of 1 in MDI, which is equivalent to the special case of probability matching for this policy (i.e., choosing arms in the same proportion as the estimated reward rates). Thus, those MDI who are presumably good learners (i.e., using a DBM learning-based strategy) showed lower preference toward options they estimated to have the highest pay-off rate. Further consistent with this reward hyposensitivity, MDI as a group appears to have a lower prior expectation of reward in their environment (10% reward rate) relative to controls (40%).

These findings are congruent with evidence of decreased incentive salience and altered reward sensitivity toward non-drug rewards observed in substance users (Chang et al., 2002). For instance, relative to control subjects, cocaine-addicted individuals showed reduced activation of the left OFC for high gains in a forced-choice task under three performance-based monetary reward conditions (Goldstein and Volkow, 2011). In this study, cocaine abusers were also less sensitive to differences between monetary rewards in left OFC and in DLPFC, with a majority exhibiting flat value ratings of all monetary amounts received in the task. Similarly, a recent study found that MDI had weaker neural responses to the anticipation of a pleasant interoceptive stimulus with mechano-receptive C-fiber stimulation (i.e., forearm and palm pleasant touch), which was apparent in the anterior insula, dorsal striatum, and thalamus (Goldstein et al., 2007). Together with the present findings, this research points to a decreased sensitivity to non-drug rewards in MDI, which has been linked to negative emotionality and thus may pause a challenge for the therapeutic rehabilitation of these patients (Koob and Le Moal, 2001; Goldstein and Volkow, 2011; May et al., 2013). Thus, future studies should examine how stimulant and other substance-dependent individuals respond to non-drug related reinforcers. A computational approach, as shown here, may be particularly useful to tease apart subtle reward sensitivity and strategic alterations based on behavioral data, without the use of cost-heavy methods such as fMRI.

Finally, using VBM analysis of structural brain scans, we found a positive correlation between participant's Softmax reward maximization parameter and gray matter volume of the thalamic LD nucleus. Importantly, MDI also had lower LD gray matter volumes relative to HCS, and this relationship was mediated by the Softmax reward maximization parameter. The lateral dorsal nucleus is part of the limbic system and has been implicated in emotional processing. It receives input from the hippocampal gyrus (Leventhal et al., 2010) and is connected reciprocally with the cingulate gyrus, a region involved in decision-making and the processing of conflict and expectancy violation (Somerville et al., 2006; Alberstone, 2009). Moreover, through its connections to the retrosplenial area as well as the pre— and parasubiculum, the LD is thought to be involved in the integration of directional information for spatial navigation (Kennerley et al., 2011) and may contribute to supporting episodic memory and mental imagery of future events (Kaitz and Robertson, 1981; Maguire et al., 2003). Thus, while these data are preliminary and require confirmation in larger samples, our findings suggest that this thalamic structure could also play an important role in modulating reward sensitivity and choice behavior during sequential goal-directed decision-making.

SUMMARY

Using a computational approach, we found evidence of cognitive abnormalities underlying reward-based learning and decision-making in MDI. Such alterations were apparent at several levels, including beliefs about hidden reward rates in the environment (MDI had lower prior expectations of reward), the type of strategy/decision policy used (more MDI relied on a myopic learning-independent strategy), and the extent of bias toward choosing the options believed to have the highest reward rates (MDI exhibited lower reward maximization bias based on the Softmax policy).

Given the absence of group differences on coarse behavioral measures, such as reaction times and points earned in the game, our results suggests that a sophisticated computational modeling approach can be a powerful neuropsychological tool to capture a combination of subtle learning and strategic abnormalities in clinical populations. Therefore, such models could be particularly useful to more precisely and comprehensively identify behavioral and neurological markers of cognitive deficits in substance-using individuals, which in turn may help develop better clinical risk prediction models. It will be critical in future research to identify how these computationally derived cognitive biases can predict the development and maintenance of the addiction cycle, including the recurrence of cravings and drug seeking behavior. For instance, while all cognitive abnormalities identified here are likely to partly contribute to behavioral dysfunction in MDI, a decreased responsiveness to non-drug rewarding stimuli may play a prominent role in perpetuating anhedonia and negative emotionality, which in turn may lead to craving and increased likelihood of relapse. Behavioral interventions aimed at boosting healthy reward responsiveness and positive affect are thus worth investigating as possible tools for promoting substance abuse prevention and recovery.

FUNDING

This study was funded by NIDA grant F32DA036959 (to KH).

REFERENCES

Alberstone, C. D. (2009). *Anatomic Basis of Neurologic Diagnosis.* Stuttgart: Thieme.

Andersson, J. L. R., Jenkinson, M., and Smith, S. (2007). *Non-linear Registration, aka Spatial Normalisation.* FMRIB Technical Report TR07JA2. FMRIB Analysis Group of the University of Oxford, Oxford.

Aron, J. L., and Paulus, M. P. (2007). Location, location: using functional magnetic resonance imaging to pinpoint brain differences relevant to stimulant use. *Addiction* 102, 33–43. doi: 10.1111/j.1360-0443.2006.01778.x

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Ball, T. M., Stein, M. B., Ramsawh, H. J., Campbell-Sills, L., and Paulus, M. P. (2014). Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology* 39, 1254–1261. doi: 10.1038/npp.2013.328

Baron, R. M., and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173. doi: 10.1037/0022-3514.51.6.1173

Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221. doi: 10.1038/nn1954

Brecht, M.-L., and Herbeck, D. (2014). Time to relapse following treatment for methamphetamine use: a long-term perspective on patterns and predictors. *Drug Alcohol Depend*. 139, 18–25. doi: 10.1016/j.drugalcdep.2014.02.702

Bunge, S. A., Wallis, J. D., Parker, A., Brass, M., Crone, E. A., Hoshi, E., et al. (2005). Neural circuitry underlying rule use in humans and nonhuman primates. *J. Neurosci.* 25, 10347–10350. doi: 10.1523/JNEUROSCI.2937-05.2005

Buxton, J. A., and Dove, N. A. (2008). The burden and management of crystal meth use. *Can. Med. Assoc. J.* 178, 1537–1539. doi: 10.1503/cmaj.071234

Chang, L., Ernst, T., Speck, O., Patel, H., Desilva, M., Leonido-Yee, M., et al. (2002). Perfusion MRI and computerized cognitive test abnormalities in abstinent methamphetamine users. *Psychiatry Res.* 114, 65–79. doi: 10.1016/S0925-4927(02)00004-5

Clark, V. P., Beatty, G. K., Anderson, R. E., Kodituwakku, P., Phillips, J. P., Lane, T. D. R., et al. (2012). Reduced fMRI activity predicts relapse in patients recovering from stimulant dependence. *Hum. Brain Mapp*. 35, 414–428. doi: 10.1002/hbm.22184

Cohen, J. D., McClure, S. M., and Angela, J. Y. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. B* 362, 933–942. doi: 10.1098/rstb.2007.2098

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879. doi: 10.1038/nature04766

Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., et al. (2007). Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain* 130, 2375–2386. doi: 10.1093/brain/awm184

Erev, I., Ert, E., and Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *J. Behav. Decis. Mak.* 21, 575–597. doi: 10.1002/bdm.602

Ersche, K., Fletcher, P., Lewis, S. J., Clark, L., Stocks-Gee, G., London, M., et al. (2005). Abnormal frontal activations related to decision-making in current and former amphetamine and opiate dependent individuals. *Psychopharmacology* 180, 612–623. doi: 10.1007/s00213-005-2205-7

Fonzo, G. A., Flagan, T. M., Sullivan, S., Allard, C. B., Grimes, E. M., Simmons, A. N., et al. (2013). Neural functional and structural correlates of childhood maltreatment in women with intimate-partner violence-related posttraumatic stress disorder. *Psychiatry Res.* 211, 93–103. doi: 10.1016/j.pscychresns.2012.08.006

Frazier, P. I., Powell, W. B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM J. Cont. Optim.* 47, 2410–2439. doi: 10.1137/070693424

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 457–472. doi: 10.1214/ss/1177011136

Good, C. D., Ashburner, J., and Frackowiak, R. S. (2001). Computational neuroanatomy: new perspectives for neuroradiology. *Rev. Neurol. (Paris).* 157(8-9 Pt 1):797–806.

Goldstein, R. Z., Alia-Klein, N., Tomasi, D., Zhang, L., Cottone, L. A., Maloney, T., et al. (2007). Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction? *Am. J. Psychiatry* 164, 43–51. doi: 10.1176/ajp.2007.164.1.43

Goldstein, R. Z., and Volkow, N. D. (2011). Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nat. Rev. Neurosci.* 12, 652–669. doi: 10.1038/nrn3119

Gonzalez, C., and Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychol. Rev.* 118, 523. doi: 10.1037/a0024558

Gowin, J. L., Harlé, K. M., Stewart, J. L., Wittmann, M., Tapert, S. F., and Paulus, M. P. (2014). Attenuated insular processing during risk predicts relapse in early abstinent methamphetamine-dependent individuals. *Neuropsychopharmacology* 39, 1379–1387. doi: 10.1038/npp.2013.333

Harlé, K. M., Shenoy, P., Stewart, J. L., Tapert, S. F., Yu, A. J., and Paulus, M. P. (2014). Altered neural processing of the need to stop in young adults at risk for stimulant dependence. *J. Neurosci.* 34, 4567–4580. doi: 10.1523/JNEUROSCI.2297-13.2014

Hesselbrock, M., Easton, C., Bucholz, K. K., Schuckit, M., and Hesselbrock, V. (1999). A validity study of the SSAGA–a comparison with the SCAN. *Addiction* 94, 1361–1370. doi: 10.1046/j.1360-0443.1999.94913618.x

Hills, T. T., and Hertwig, R. (2012). Two distinct exploratory behaviors in decisions from experience: comment on Gonzalez and Dutt (2011). *Psychol. Rev.* 119, 888–892. doi: 10.1037/a0028004

Hoffman, W. F., Moore, M., Templin, R., McFarland, B., Hitzemann, R. J., and Mitchell, S. H. (2006). Neuropsychological function and delay discounting in methamphetamine-dependent individuals. *Psychopharmacology* 188, 162–170. doi: 10.1007/s00213-006-0494-0

Ide, J. S., Shenoy, P., Yu, A. J., and Li, C. S. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. *J. Neurosci.* 33, 2039–2047. doi: 10.1523/JNEUROSCI.2201-12.2013

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *J. Artif. Intell. Res.* 4, 237–285. doi: 10.1613/jair.301

Kaitz, S. S., and Robertson, R. T. (1981). Thalamic connections with limbic cortex. II. Corticothalamic projections. *J. Compar. Neurol.* 195, 527–545. doi: 10.1002/cne.901950309

Kennerley, S. W., Behrens, T. E., and Wallis, J. D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nat. Neurosci.* 14, 1581–1589. doi: 10.1038/nn.2961

Koob, G. F., and Le Moal, M. (2001). Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology* 24, 97–129. doi: 10.1016/S0893-133X(00)00195-0

Lee, M. D., Zhang, S., Munro, M., and Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cogn. Syst. Res.* 12, 164–174. doi: 10.1016/j.cogsys.2010.07.007

Leland, D. S., and Paulus, M. P. (2005). Increased risk-taking decision-making but not altered response to punishment in stimulant-using young adults. *Drug Alcohol Depend.* 78, 83–90. doi: 10.1016/j.drugalcdep.2004.10.001

Leventhal, A. M., Brightman, M., Ameringer, K. J., Greenberg, J., Mickens, L., Ray, L. A., et al. (2010). Anhedonia associated with stimulant use and dependence in a population-based sample of American adults. *Exp. Clin. Psychopharmacol.* 18, 562. doi: 10.1037/a0021964

London, E. D., Simon, S. L., Berman, S. M., Mandelkern, M. A., Lichtman, A. M., Bramen, J., et al. (2004). Mood disturbances and regional cerebral metabolic abnormalities in recently abstinent methamphetamine abusers. *Arch. Gen. Psychiatry* 61:73. doi: 10.1001/archpsyc.61.1.73

Luce, R. (1959). *Individual Choice Behavior.* New York, NY: Wiley.

Ma, N., and Yu, A. J. (2015). Statistical learning and adaptive decision-making underlie human response time variability in inhibitory control. *Front. Psychol.* 6:1046. doi: 10.3389/fpsyg.2015.01046

Maguire, E. A., Valentine, E. R., Wilding, J. M., and Kapur, N. (2003). Routes to remembering: the brains behind superior memory. *Nat. Neurosci.* 6, 90–95. doi: 10.1038/nn988

May, A. C., Stewart, J. L., Migliorini, R., Tapert, S. F., and Paulus, M. P. (2013). Methamphetamine dependent individuals show attenuated brain response to pleasant interoceptive stimuli. *Drug Alcohol Depend.* 131, 238–246. doi: 10.1016/j.drugalcdep.2013.05.029

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Ann. Rev. Neurosci.* 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Monterosso, J. R., Ainslie, G., Xu, J., Cordova, X., Domier, C. P., and London, E. D. (2007). Frontoparietal cortical activity of methamphetamine—dependent and comparison subjects performing a delay discounting task. *Hum. Brain Mapp.* 28, 383–393. doi: 10.1002/hbm.20281

Panenka, W. J., Procyshyn, R. M., Lecomte, T., MacEwan, G. W., Flynn, S. W., Honer, W. G., et al. (2013). Methamphetamine use: a comprehensive review of molecular, preclinical and clinical findings. *Drug Alcohol Depend.* 129, 167–179. doi: 10.1016/j.drugalcdep.2012.11.016

Paulus, M. P., Hozack, N. E., Zauscher, B. E., Frank, L., Brown, G. G., Braff, D. L., et al. (2002). Behavioral and functional neuroimaging evidence for prefrontal dysfunction in methamphetamine-dependent subjects. *Neuropsychopharmacology* 26, 53–63. doi: 10.1016/S0893-133X(01)00334-7

Paulus, M. P., Hozack, N., Frank, L., Brown, G. G., and Schuckit, M. A. (2003). Decision making by methamphetamine-dependent subjects is associated with error-rate-independent decrease in prefrontal and parietal activation. *Biol. Psychiatry* 53, 65–74. doi: 10.1016/S0006-3223(02)01442-7

Paulus, M. P., and Stein, M. B. (2006). An insular view of anxiety. *Biol. Psychiatry* 60, 383–387. doi: 10.1016/j.biopsych.2006.03.042

Paulus, M. P., Tapert, S. F., and Schuckit, M. A. (2005). Neural activation patterns of methamphetamine-dependent subjects during decision making predict relapse. *Arch. Gen. Psychiatry* 62:761. doi: 10.1001/archpsyc.62.7.761

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* 58, 527–535. doi: 10.1090/S0002-9904-1952-09620-8

Rogers, R. D., Everitt, B., Baldacchino, A., Blackshaw, A., Swainson, R., Wynne, K., et al. (1999). Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: evidence for monoaminergic mechanisms. *Neuropsychopharmacology* 20, 322–339. doi: 10.1016/S0893-133X(98)00091-8

Ryzhov, I. O., Powell, W. B., and Frazier, P. I. (2012). The knowledge gradient algorithm for a general class of online learning problems. *Oper. Res.* 60, 180–195. doi: 10.1287/opre.1110.0999

Scorzin, J. E., Kaaden, S., Quesada, C. M., Müller, C.-A., Fimmers, R., Urbach, H., et al. (2008). Volume determination of amygdala and hippocampus at 1.5 and 3.0 T MRI in temporal lobe epilepsy. *Epilepsy Res.* 82, 29–37. doi: 10.1016/j.eplepsyres.2008.06.012

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23, S208–S219. doi: 10.1016/j.neuroimage.2004.07.051

Somerville, L. H., Heatherton, T. F., and Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nat. Neurosci.* 9, 1007–1008. doi: 10.1038/nn1728

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (2003). *WinBUGS Version 1.4. Bayesian Inference using Gibbs Sampling.* Cambridge: MRC Biostatistics Unit, Institute for Public Health.

Steyvers, M., Lee, M. D., and Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *J. Math. Psychol.* 53, 168–179. doi: 10.1016/j.jmp.2008.11.002

Tabibnia, G., Monterosso, J. R., Baicy, K., Aron, A. R., Poldrack, R. A., Chakrapani, S., et al. (2011). Different forms of self-control share a neurocognitive substrate. *J. Neurosci.* 31, 4805–4810. doi: 10.1523/JNEUROSCI.2859-10.2011

Uttl, B. (2002). North American adult reading test: age norms, reliability, and validity. *J. Clin. Exp. Neuropsychol.* 24, 1123–1137. doi: 10.1076/jcen.24.8.1123.8375

van der Plas, E. A., Crone, E. A., van den Wildenberg, W. P., Tranel, D., and Bechara, A. (2009). Executive control deficits in substance-dependent individuals: a comparison of alcohol, cocaine, and methamphetamine and of men and women. *J. Clin. Exp. Neuropsychol.* 31, 706–719. doi: 10.1080/13803390802484797

Yu, A., and Cohen, J. (2009). Sequential effects: superstition or rational behavior. *Adv. Neural Inf. Process. Syst.* 21, 1873–1880.

Yu, A. J., and Huang, H. (2014). Maximizing masquerading as matching in human visual search choice behavior. *Decision* 1, 275. doi: 10.1037/dec0000013

Zhang, S., Huang, C. H., and Angela, J. Y. (2014). "Sequential effects: a Bayesian analysis of prior bias on reaction time and behavioral choice," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (Québec City, QC).

Zhang, S., and Yu, J. A. (2013a). Cheap but clever: human active learning in a bandit setting. *Ratio* 12:14. Available online at: http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0306/paper0306.pdf

Zhang, S., and Yu, J. A. (2013b). "Forgetful Bayes and myopic planning: human learning and decision-making in a bandit setting," in *Advances in Neural Information Processing Systems* (Lake Tahoe), 2607–2615.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

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