frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 11 February 2014
doi: 10.3389/fpsyg.2014.00091

Emergent structured transition from variation to repetition in a biologically-plausible model of learning in basal ganglia

Ashvin Shah* and Kevin N. Gurney

Department of Psychology, University of Sheffield, Sheffield, UK

Edited by:
Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:
Fred H. Hamker, Chemnitz University of Technology, Germany
Vincenzo G. Fiore, UCL Institute of Neurology, UK
Valerio Sperati, Consiglio Nazionale delle Ricerche, Italy

****Correspondence:***
Ashvin Shah, Department of Psychology, University of Sheffield, Western Bank, Sheffield S10 2TP, UK
e-mail: a.shah@sheffield.ac.uk

Often, when animals encounter an unexpected sensory event, they transition from executing a variety of movements to repeating the movement(s) that may have caused the event. According to a recent theory of action discovery (Redgrave and Gurney, 2006), repetition allows the animal to represent those movements, and the outcome, as an action for later recruitment. The transition from variation to repetition often follows a non-random, structured, pattern. While the structure of the pattern can be explained by sophisticated cognitive mechanisms, simpler mechanisms based on dopaminergic modulation of basal ganglia (BG) activity are thought to underlie action discovery (Redgrave and Gurney, 2006). In this paper we ask the question: can simple BG-mediated mechanisms account for a structured transition from variation to repetition, or are more sophisticated cognitive mechanisms always necessary? To address this question, we present a computational model of BG-mediated biasing of behavior. In our model, unlike most other models of BG function, the BG biases behavior through modulation of cortical response to excitation; many possible movements are represented by the cortical area; and excitation to the cortical area is topographically-organized. We subject the model to simple reaching tasks, inspired by behavioral studies, in which a location to which to reach must be selected. Locations within a target area elicit a reinforcement signal. A structured transition from variation to repetition emerges from simple BG-mediated biasing of cortical response to excitation. We show how the structured pattern influences behavior in simple and complicated tasks. We also present analyses that describe the structured transition from variation to repetition due to BG-mediated biasing and from biasing that would be expected from a type of cognitive biasing, allowing us to compare behavior resulting from these types of biasing and make connections with future behavioral experiments.

Keywords: action discovery, reinforcement, basal ganglia, variation, repetition

1. INTRODUCTION

Animals are capable of executing a huge variety of movements but, importantly, they can discover the specific movements that affect the environment in predictable ways and represent them as *actions* for later recruitment. Redgrave, Gurney, and colleagues have suggested that this occurs through a process they refer to as *action discovery* (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013). Action discovery begins when an animal is executing movements within some context and an unexpected salient sensory event (such as a light flash) occurs. The unexpected sensory event causes a short-latency phasic increase in dopamine (DA) neuron activity (henceforth referred to simply as *DA activity*). Through its influence on the basal ganglia (BG)—a group of interconnected subcortical structures which, in turn, influence cortical activity—the increase in DA activity can help bias the animal to repeat the movements that preceded the unexpected sensory event under the same contextual circumstances. This *repetition bias* (Redgrave and Gurney, 2006) allows associative networks in the brain to learn and encode the movements as an action because it causes a frequent and reliable presentation of context, movements, and the sensory event as the outcome of those movements.

This transition from executing a variety of movements to repeating just one or a subset of movements often follows a non-random, structured, pattern. For example, consider a spatial task such that reaching to a specific location results in the outcome. Here, one type of structured transition from variation to repetition occurs if the animal gradually refines its movements so that movements that are further from the location decrease in frequency earlier than movements that are closer to the location.

The non-random structure of the transition from variation to repetition can be explained with "intelligent" or sophisticated cognitive mechanisms, e.g., by using an estimation of the range of movements that cause the outcome that gets more and more precise with repeated occurrences of the outcome. Similarly, other types of a structured transition may rely on other sophisticated notions such as optimality or uncertainty (e.g., Dearden et al. 1998; Dimitrakakis 2006; Simsek and Barto 2006). However, the process of action discovery is thought to be mediated primarily by simpler mechanisms involving DA modulation of the BG, and not sophisticated cognitive mechanisms. In this paper we ask the question, can *simple* BG-mediated mechanisms guide a structured transition from variation to repetition, or must sophisticated cognitive mechanisms always be recruited? To address this question, we present a computational model of BG-mediated biasing of behavior.

Our model will necessarily deal with a specific and, therefore, limited example of action discovery and so to establish its status, we now outline the model's wider context comprising various broad categories of action. For example, one type of action might involve making a particular gesture with the hand (as in sign language or hand signaling), regardless of the precise spatial location of the hand, and no environmental object is targeted. Another type of action involves manipulating objects in the environment (such as flipping a light switch or typing out a password). In this instance, space is weakly implicit (the objects are located somewhere); the key feature is the target object identity and its manipulation. In this paper, we focus on an explicitly spatial task: the relatively simple action of moving an end-effector to a particular spatial location. In the model task, a movement end-point to which to move must be selected. End-points that correspond to a target location elicit a reinforcement signal, and, importantly, reinforcement is not contingent on movement trajectory. The model task is inspired by behavioral counterparts we have used to study action discovery in which participants manipulate a joystick to find an invisible target area in the workspace (Stafford et al., 2012, 2013; Thirkettle et al., 2013a,b). While there may be "gestural" aspects of action in the behavioral task, in the model we ignore these and focus only on the spatial location of movement end-point.

In the next few paragraphs, we describe features of neural processing which our model incorporates that many other models of the BG do not. Biological theories of BG function suggest that the BG bias behavior not through direct excitation of their efferent targets, but, rather, through the selective relaxation of inhibition (i.e., disinhibition) of their efferent targets (Chevalier and Deniau, 1990; Mink, 1996; Redgrave et al., 2011). When the BG are presented with multiple signals, each representing an action or movement, these signals will have different activity levels signifying the urgency or *salience* of the "action request." BG are supposed to process each signal through a neural population or *channel*, and inter-channel connections facilitate competitive processes resulting in suppression of BG output (inhibition) on high salience channels and increased output on the low salience channels (Gurney et al., 2001a,b; Humphries and Gurney, 2002; Prescott et al., 2006). Many models of BG function focus on how the multiple signals presented to the BG are transformed to the activity of the BG's output nucleus. Action selection in these models is then based on the latter's activity (e.g., Gurney et al. 2001a,b, 2004; Joel et al. 2002; Daw et al. 2005; Shah and Barto 2009). However, one important feature of our model is that it also takes into account the pattern of excitation from other areas to the BG's efferent targets (see also Humphries and Gurney 2002; Cohen and Frank 2009; Baldassarre et al. 2013). Thus, behavior results from BG modulation of their efferent target's response to excitation patterns, and is not just a mirror of the activity of the BG's output nucleus.

Further, many models of BG function focus on how the BG select from a small number of abstract independent behaviors (e.g., Gurney et al. 2001b; Daw et al. 2005; Cohen and Frank 2009; Shah and Barto 2009). While such representations may be appropriate for some behavioral tasks in experimental psychology, in ethological action discovery, the space of activities from which to select may be larger and adhere to some inherent topology. In our model, candidate locations to which to move are represented by a large number of topographically-organized neurons in cortex so that neighboring spatial locations are represented by neighboring neurons. Excitation to cortex follows a pattern in which all neurons are weakly excited initially, and that pattern evolves so that eventually only one neuron is excited strongly. This pattern is inspired by neural activity observed in perceptual decision-making tasks (Britten et al., 1992; Platt and Glimcher, 1999; Huk and Shadlen, 2005; Gold and Shadlen, 2007), and as suggested by evidence accumulation models of decision-making (Bogacz et al., 2006; Lepora et al., 2012).

We hypothesize that because the BG bias behavior by modulating cortical response to excitation, and that that excitation follows a structured pattern, simple BG-mediated biasing can result in a structured transition from variation to repetition in action discovery. Sophisticated cognitive mechanisms are not necessarily required to develop a structured transition.

In addition, behavioral biasing in action discovery is not thought to be driven by "extrinsic motivations" that are based on rewarding consequences and that dictate reinforcement in many types of operant conditioning tasks (Thorndike, 1911; Skinner, 1938) and computational reinforcement learning (RL) (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). Rather, "intrinsic motivations" (Oudeyer and Kaplan, 2007; Baldassarre, 2011; Barto, 2013; Barto et al., 2013; Gottlieb et al., 2013; Gurney et al., 2013) that are triggered by the occurrence of an unexpected sensory event may drive DA activity and thus behavioral biasing in action discovery (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013; Mirolli et al., 2013). In such cases, if the outcome does not represent or predict an extrinsically-rewarding event, reinforcement decreases as associative networks in the brain learn to predict its occurrence (Redgrave and Gurney, 2006; Redgrave et al., 2011). Rather than implement a model of prediction explicitly, we approximate its effects with a simple model of habituation in which the rate of reinforcement decreases as the target location is repeatedly hit (Marsland, 2009). This habituation model approximates the dependence of DA activity on outcome predictability in action discovery (Redgrave and Gurney, 2006; Redgrave et al., 2011), and is similar to that used in neural network models of novelty detection (Marsland, 2009).

In this paper, we use computational models to demonstrate that simple BG-mediated mechanisms can bias behavior, via their modulation of cortical response to a pattern of excitation, such that the transition from variation to repetition follows a structured pattern. We describe this structured pattern and show how it, along with the effects of habituation, lead to behavioral patterns in tasks in which one target area delivers a reinforcement signal, two target areas deliver reinforcement, or the target area that delivers reinforcement changes location. These experiments lead to predictions as to the type of behavior that would be expected when only simple BG-mediated mechanisms, and not more sophisticated cognitive mechanisms, bias behavior. We also run models that mimic a simple form of transition from variation to repetition that would be expected under sophisticated cognitive mechanisms by subsuming the effects of those mechanisms in a phenomenological way. In order to make contact with future behavioral experiments, we develop a novel characterization of behavioral trends which links these trends to underlying neural mechanisms that dictate different forms of biasing.

2. METHODS

We use a computational model, based on established models (Gurney et al., 2001a,b; Humphries and Gurney, 2002), to control movement selection in a task that simulates reaching or pointing to specific target spatial locations. We provide here a conceptual overview of its mechanics; detailed equations are provided in the Supplementary section.

The model is a neural network model with leaky-integrator neuron units (henceforth referred to as "neurons" for brevity), the activities of which represent conglomerate neural firing rate of a group of neurons (Gurney et al., 2001a,b). Each brain area in the model, except for the area labeled "Context," consists of 196 neurons spatially arranged in a 14 × 14 grid. Each neuron in each area is part of an "action channel" (Gurney et al., 2001a,b; Humphries and Gurney, 2002) such that its location in the grid corresponds to a movement toward the corresponding location of a two-dimensional workspace. For the purposes of this model, the workspace is of dimensions 14 × 14 units. Most projections from one area to another are one-to-one and not plastic; exceptions will be explicitly noted.

Figure 1 illustrates the gross architecture of the model. In brief, the end-point location of a movement, X_M, is determined by the activities of neurons in "M (Cortex)." These neurons are excited by an exploratory mechanism, "E (Explorer)," and are engaged in positive feedback loops with neurons in "T (Thalamus)." The basal ganglia (BG, gray boxes) send inhibitory projections to Thalamus neurons, and they modulate the gain of the Cortex-Thalamus positive feedback loops (Chambers et al., 2011) through selective disinhibition of Thalamus neurons. Cortex and Thalamus represent grids of neurons that correspond to motor-related areas of cortex and thalamus, respectively.

2.1. EXCITATORY INPUTS TO THE NEURAL NETWORK

There are two sources of excitatory input to the neural network.The first is labeled "C (Context)" and represents the context, such as participating in the current experiment. There is only one context for the results reported in this paper. Thus, Context consists of a single neuron with an output activity set to a constant value. Context influences BG activity through one-to-all projections to areas D1, D2, and STN. Projections to D1 and D2 are plastic and represent a context-dependent biasing of movements, as described in the subsection "Biasing of behavior."

The second source of excitatory input is "E (Explorer)," which provides excitation to Cortex which, in turn, is responsible for movement. The Explorer is the source of variation required to explore the space of possible movements. This variation may be more or less random or structured according to the strategy used. However, these strategies are devised by other mechanisms, not explicitly modeled here, and we simply aim to capture the effects of such strategies in the Explorer.

In this paper, the Explorer is inspired by a range of experimental data. First, recordings in some areas of parietal cortices (Anderson and Buneo, 2002) show activation of neurons corresponding to a decision to make a movement that terminates at the location represented by those neurons. Further, several experimental studies, (Britten et al., 1992; Platt and Glimcher, 1999; Huk and Shadlen, 2005; Gold and Shadlen, 2007) show that neurons representing different decisions are weakly active early in the decision-making process. The activities of some neurons—corresponding to the executed decision in these experiments—increase at a greater rate than that of other neurons.

We capture features of this behavior with a hand-crafted function describing, for a decision to move to a particular spatial location, the evolution of activity for every neuron in the Explorer. Early in the process, all neurons are weakly-excited with low activation levels. Neural activity evolves such that, as confidence in a particular movement increases, so does the corresponding neuron activity. The activities of other neurons increase to a lesser degree. An example of this behavior is shown in **Figure 2**; it is described in greater detail in the next paragraph and in the Supplementary section.

For each movement, a particular neuron in Explorer, labeled G_{exp}, is chosen. If we suppose that sophisticated cognitive mechanisms are not devoted to movement selection, G_{exp} is chosen randomly. The activity of the neuron corresponding to G_{exp} increases linearly to one (green line in **Figure 2**). The activities of surrounding neurons change according to a Gaussian-like function centered at G_{exp}. They first increase and then decrease; those furthest from G_{exp} increase by a small amount and then quickly decrease to zero, while those closer to G_{exp} increase by a larger amount and decrease at a later time point to zero. The pattern of activity such that the activity of neuron G_{exp} is one and the activities of all other neurons are at zero is held for brief time, and then the activities of all neurons are set to zero. This evolution takes T_E time steps, which is the number of time steps in a trial.

If, in contrast, we assume sophisticated cognitive mechanisms do influence movement selection, G_{exp} is chosen in order to reflect that strategy, e.g., according to some heuristic search such as a spiral pattern or quadrant-by-quadrant search. In this paper we examine behavior that results when cognitive mechanisms do not influence movement selection as well as behavior that results from a simple pattern, as described in the subsection "Biasing of behavior."

2.2. CORTEX AND THALAMUS

"Cortex" represents cortical areas that encode high-level movement plans such as reaching or pointing to a location (Anderson and Buneo, 2002). In our model, the spatial location of a neuron in Cortex corresponds to a target spatial location in the workspace, or movement end-point, to which to reach. Cortex (M) receives excitatory projections from Explorer and

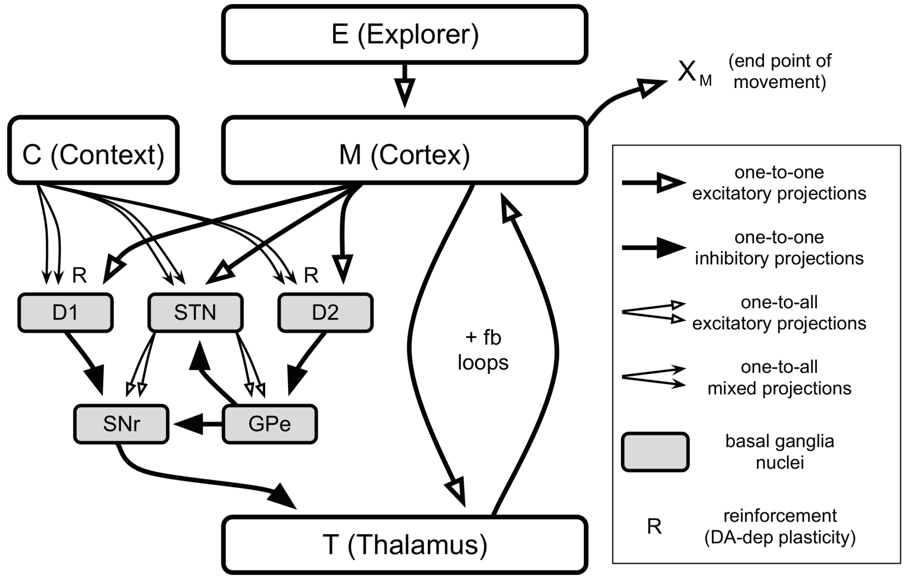
E (Explorer)
X_M (end point of movement)
C (Context)
M (Cortex)
R
D1
STN
D2
R
+ fb loops
SNr
GPe
T (Thalamus)
one-to-one excitatory projections
one-to-one inhibitory projections
one-to-all excitatory projections
one-to-all mixed projections
basal ganglia nuclei
R
reinforcement (DA-dep plasticity)

FIGURE 1 | Architecture of the model. Each box except for "C (Context)" contains 196 neurons spatially arranged in a 14 × 14 grid. Context contains just one neuron. Types of projections are labeled in the legend on the right.

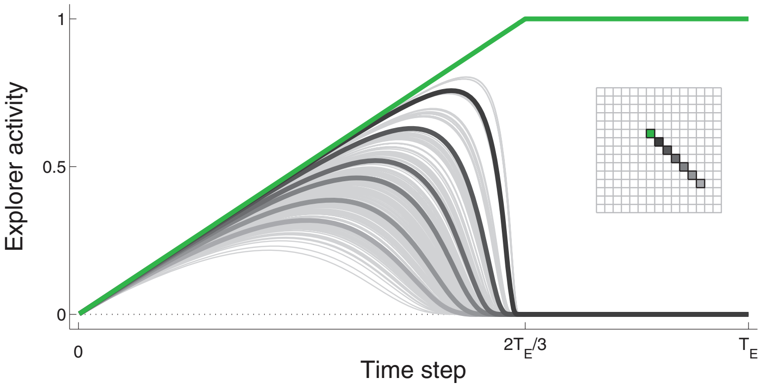

FIGURE 2 | Example of activity of Explorer neurons during a typical movement. The activity of the neuron corresponding to the focus of excitation, G_{exp}, is drawn in green. Selected neurons, colored in the inset, are drawn with thick lines in different shades of gray so as to demonstrate the spatial influence on excitation pattern. All other neurons are drawn in thin gray lines

Thalamus (T) which preserve channel identity; that is, the neurons representing a given channel in Explorer and Thalamus project to the corresponding neuron in Cortex. In turn, Thalamus receives channel-wise excitatory projections from Cortex, and channel-wise inhibitory projections from SNr (a nucleus of the BG called the substantia nigra pars reticulata). Cortex and Thalamus therefore form a positive feedback loop referred to as a *Cortex-Thalamus loop*, for each channel which is excited by the corresponding channel in Explorer. The gain of a Cortex-Thalamus loop is modulated by inhibitory projections from SNr neuron to Thalamus (Chambers et al., 2011). When the activity level of an SNr channel is low, the corresponding Thalamus neuron is said to be *disinhibited* and its Cortex-Thalamus loop has a high gain. A Cortex-Thalamus loop with a high gain is more easily-excited by the corresponding Explorer neuron.

2.3. BASAL GANGLIA

The functional properties of BG architecture have been described in detail in prior work (Gurney et al., 2001a,b; Humphries and Gurney, 2002; Redgrave et al., 2011). Briefly, the BG is a subcortical group of brain areas with intrinsic architecture that is well-suited to select one behavioral option among competing options. The BG implement an off-center on-surround excitation pattern: The BG channel i that is most strongly-excited by its cortical "action request" inhibits the corresponding target channel (neuron) in Thalamus the least, while other Thalamus channels $j \neq i$ are further inhibited. Thus, Cortex-Thalamus loop i is most easily-excited by input from Explorer to Cortex, and other Cortex-Thalamus loops $j \neq i$ are harder to excite by input from Explorer to Cortex. These properties are similar in some ways to those of a winner-take-all network between the competing channels, but additional architectural features of the BG ensure better control of the balance between excitation and inhibition (Gurney et al., 2001a,b). D1 and D2 refer to different populations of neurons (named after the dopamine receptors they predominantly-express) in a nucleus of the BG called the striatum. The pathway comprising D1 and STN (subthalamic nucleus) performs the selection with an off-center on-surround network in which D1 supplies focussed ("central") inhibition and the STN a diffuse ("surround") excitation. The pathway through D2 regulates the selection by controlling, though GPe (external segment of the globus pallidus), the excitatory activity of STN (Gurney et al., 2001a,b).

2.4. FROM CORTICAL ACTIVITY TO BEHAVIOR

Movement in this model is a function of the activities of the Cortex neurons. Each neuron with an activation greater than a threshold η "votes" to move to the location represented by its grid location with a strength proportional to its activity (i.e., using a population code, Georgopoulos et al. 1982). In most cases, because of the selection properties of the BG, the activation of only one Cortex neuron rises above η. At each time step t, the target location to which to move, $X_M(t)$, is an average of the locations represented by Cortex neurons with activities above η, weighted by their activities. At each t, if any Cortex neuron is above η, a simple "motor plant" causes a movement from the current position ($x_p(t)$) toward $X_M(t)$ (see Supplementary section for equations). Movement evaluation, and hence any learning, is based only on $x_p(T_E)$, the position at time T_E (the last time step of a trial). Thus, end-point of movement, not movement trajectory, is evaluated in this model.

2.5. BIASING OF BEHAVIOR

Targets are circular areas within the workspace. A target is considered hit when $||x_p(T_E) - X_G|| < \theta_G$, where X_G is the location of the center of target G and θ_G ($= 1.1$) is the radius. Thus, a movement to the location represented by neuron i that corresponds to the center of the target, or to locations represented by the immediate four neighboring neurons, is within the target's radius. When a target is hit, behavior is biased so that the model is more likely to make movements to the target. This repetition bias (Redgrave and Gurney, 2006) can be implemented in two ways in this model.

The first way is "BG-mediated biasing," which is based on dopamine-dependent plasticity at the corticostriatal synapses (Calabresi et al., 2007; Wickens, 2009), and is implemented as a Hebbian-like rule governing plasticity to weights onto striatal D1 and D2 neurons. When the end-point of movement is evaluated (at time T_E of a trial), usually only one neuron (i) in each of Cortex, D1, and D2 have an activity above zero. If the target is hit, the weights from Cortex neuron i to D1 neuron i, Cortex neuron i to D2 neuron i, the Context neuron to D1 neuron i, and the Context neuron to D2 neuron i are increased according to equations of the following form (see Supplementary section for full equations):

$$\Delta w_i = \alpha\, \beta^{N_k - 1}\, y_{pre}\, y_{post}\, (W_{max} - w_i), \qquad (1)$$

where w_i is the weight, y_{pre} is the activity of the presynaptic neuron, y_{post} is the activity of the postsynaptic neuron, α is a step-size, $W_{\max}$ ($= 1$) is the maximum strength of a synapse, β ($= 0.825$) is a *habituation* term (Marsland, 2009), and N_k is the number of times target k has been hit. If the target is not hit, the weights are decreased. Weights from Cortex to striatum have a lower limit of zero, while weights from Context to striatum have a lower limit of -0.1. Neurons that have greater afferent weights are more-easily excited than are neurons with lower afferent weights.

Neurons in D1 and D2 that correspond to movements that were reinforced are excited by the Context neuron from the first time step of a trial onward, and neurons that correspond to movements that were not reinforced are weakly inhibited by the Context neuron. (We use negative weights to approximate the inhibitory effects of striatal interneurons, Koos and Tepper 1999; Bolam et al. 2006). Thus, weights from the Context neuron to D1 and D2 represent an *a priori* bias in favor of movements that were reinforced, and against movements that were not reinforced. This bias is context-dependent and, while there is only one context for the results reported in this paper, multiple contexts can be represented by multiple context neurons with similar learning rules. Neurons in D1 and D2 are also excited by Cortex neurons, which, early in a trial, are all weakly-excited by Explorer. Because the projections from Cortex to D1 and D2 are plastic, movements that were reinforced are more-easily excited by Cortex than movements that were not reinforced.

Thus, with BG-mediated biasing, channels corresponding to making a movement to locations that are within the target area are easily-excited by weak inputs from the Explorer after the target has been hit several times. Channels corresponding to movements that do not hit the target are made to be more difficult to excite.

The second way by which repetition bias is implemented in this model is referred to as "Cognitive biasing," whereby G_{exp} is chosen according to some strategy or pattern. Under cognitive biasing in this paper, the set of neurons in Explorer from which G_{exp} is chosen corresponds to a spatial area, centered around the location of the target, that decreases in size each time the target is hit (we describe this pattern in detail in the Supplementary section). This is a simple hand-crafted form of biasing that mimics a decrease in variation and increase in repetition by "zooming in" on the target as the target is repeatedly hit. It is meant to capture the effects of behavioral biasing as mediated by "sophisticated cognitive" or "intelligent" mechanisms. If there is no Cognitive biasing, G_{exp} is randomly chosen as described earlier.

2.6. MODEL EXPERIMENTS

A model run consists of having the model select movements for 300 trials (where a trial consists of executing one movement). Movements were reinforced (Equation 1) when they hit a particular target. We examined behavior that results from reinforcing one target, two targets simultaneously, and one target and then another. The targets are referred to G_1, $G_{2\text{far}}$ (which is far from G_1), and $G_{2\text{near}}$ (which is near G_1). Experiments 1 to 4 were conducted to describe patterns of behavior under simple, "non-intelligent," BG-mediated biasing and different conditions of reinforcement. Experiment 5 was conducted to describe patterns of behavior under BG biasing, Cognitive biasing, and both.

- **Experiment 1: Single target** (G_1): We ran 50 independent runs of 300 movements during which BG biasing (and not Cognitive biasing) was used to reinforce movements that hit G_1.
- **Experiment 2: Two simultaneous targets** (G_1 **and** $G_{2\text{far}}$): We ran 50 independent runs of 300 movements during which BG biasing was used to reinforce movements that hit either G_1 or $G_{2\text{far}}$.
- **Experiment 3: Reinforce** G_1**, then** $G_{2\text{far}}$**, then** G_1 **again**: We ran 50 independent runs of 900 movements during which BG biasing was used to reinforce movements that hit G_1 for the first 300 movements, then to reinforce movements that hit $G_{2\text{far}}$ (but not those that hit G_1) for the next 300 movements, and then reinforce movements that hit G_1 (but not those that hit $G_{2\text{far}}$) for the final 300 movements.
- **Experiment 4: Reinforce** G_1**, then either** $G_{2\text{far}}$ **or** $G_{2\text{near}}$: We ran 50 independent runs of 600 movements during which BG biasing was used to reinforce movements that hit G_1 for the first 300 movements and then to reinforce $G_{2\text{far}}$ (but not those that hit G_1) for the next 300 movements. We ran another 50 independent runs of 600 movements during which BG biasing was used to reinforce movements that hit G_1 for the first 300 movements and then to reinforce $G_{2\text{near}}$ (but not those that hit G_1) for the next 300 movements.
- **Experiment 5: Different bias conditions**: We ran 50 independent runs of 300 movements during which Cognitive biasing (and not BG biasing) was used to reinforce movements that hit G_1. We ran another 50 independent runs of 300 movements during which both BG biasing and Cognitive biasing were used to reinforce movements that hit G_1.

3. RESULTS

3.1. EXPERIMENT 1: SINGLE TARGET (G_1)

Recall that there are two sources of excitation to the model, as explained in Methods section 2.1: the Context neuron, which projects to D1, D2, and STN; and the Explorer, which projects to Cortex (see also **Figure 1**). As described in Methods section 2.1, a focus of excitation, G_{exp}, is chosen randomly, and the activities of neurons in the Explorer follow a hand-crafted pattern such that all neurons are weakly-excited initially, but that activity focuses so that only the neuron corresponding to G_{exp} is strongly-excited (see **Figure 2**). If the weights onto D1 and D2 remain at their initial values, Explorer activity will result in a movement made to the location represented by G_{exp}.

In Experiment 1, there was a single target, G_1, located in the lower right area of the work space (center of target colored in red in the upper left graph in **Figure 3**). When the target was first hit, it was because the Explorer happened to choose a G_{exp} that was within θ_G of target center. As described in Methods section 2.5, when the target is hit, the corticostriatal weights that project to striatal neurons corresponding to the movement just made are increased (Equation 1). When a target is not hit, the weights decrease. The weight change influences how the BG modulates the gain between Thalamus and Cortex positive feedback loops (Methods sections 2.2 and 2.3), and hence how Cortex responds to excitation from Explorer.

Neural activity

Figure 3 shows selected neuron activity resulting from the same excitation from the Explorer during early movements ("before learning") and during late movements ("after learning"). Excitation from Explorer is illustrated in the lower left graph, and the color scheme indicating which neuron's activity is plotted is illustrated in the upper left graph. In this example, activities of neurons corresponding movements made to G_{exp} are plotted in green; those corresponding to the center of the target (G_1) are plotted in red; and those corresponding to a subset of neurons near or between G_{exp} and G_1 are plotted in shades of gray. (Compare with **Figure 2** and Methods section 2.1.) G_{exp} is not within the target area. The top row of graphs to the right of the color scheme graph plot neuron activity in striatum D1, neuron activity in SNr, and neuron activity in Cortex in the untrained model. As excitation from Explorer evolved over time, Cortex neurons increased accordingly due to the direct one-to-one projections from Explorer to Cortex and positive feedback loops with Thalamus (as described in Methods section 2.2). Cortex activity directly excited striatal neurons due to direct one-to-one projections to striatum D1 and striatum D2 (as described in Methods section 2.3). In this case, striatal neurons corresponding to G_{exp} increased in activity. Because no learning has occurred yet, Context did not bias activity in striatum as all projections from Context to striatum remained at zero. Intra-BG processing (described in Methods section 2.3) resulted in a decrease in activity of SNr neuron corresponding to G_{exp}, and an increase in all other SNr neurons. This disinhibited the Thalamus neuron corresponding to G_{exp}, increasing the gain on the positive feedback loop with Cortex neuron corresponding to G_{exp}, thus allowing it to increase in activity even more. In addition, the increased activity of all other SNr neurons further decreased the positive feedback gain between other Cortex-Thalamus neuron pairs (Chambers et al., 2011). In this example, weights into D1 and D2 have not undergone any changes, i.e., the target has not been hit, so there is no biasing from Context. Thus, the BG facilitated the selection of the movement suggested by Explorer (move to location G_{exp}) and inhibited the selection of other movements.

After the target had been hit many times, the weights from Context to striatal neurons D1 and D2, and from Cortex to D1 and D2, that correspond to movements made to a location within the target zone (in this example, the center of G_1) increased (as described in Methods section 2.5 and Equation 1), and the weights to all others decreased by a small amount. Neuron activity in response to the same excitation from Explorer after learning is illustrated in the bottom, right most three graphs of **Figure 3**. Neurons that correspond to G_1 (plotted in red) are referred to as s_G. Because weights from Context to s_G in D1 and D2 have increased, the activity of neuron s_G in D1 and D2 increased faster due to excitation from Cortex than did that of other neurons, including that of neurons that correspond to movements made to G_{exp}. This caused a decrease in the activity of SNr neuron s_G and an increase in the gain of the corresponding Cortex-Thalamus positive feedback loop (described in Methods section 2.2). Hence, the weak excitation to Cortex neuron s_G at the beginning of a movement period was sufficient to initiate a positive feedback process between the corresponding neuron s_G in Cortex

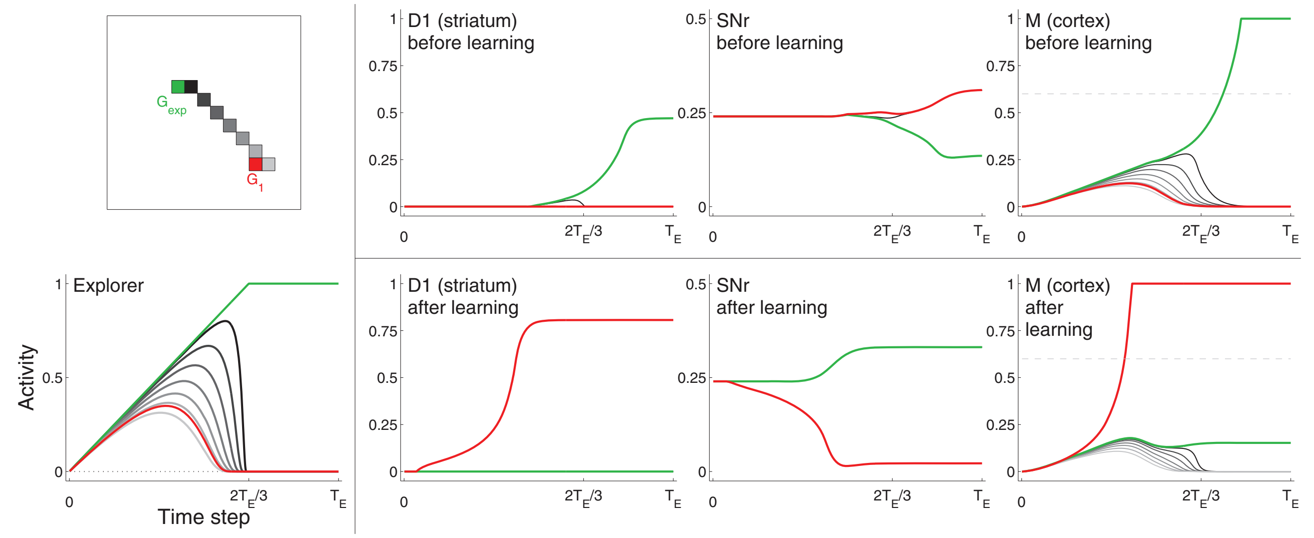

FIGURE 3 | Example neural activity for selected neurons from D1, SNr, and Cortex (M) at different points in training. Neurons are colored according to their spatial location in the grid (**top left**). The red neuron corresponds to the center of target G_1, the green neuron corresponds to the location of G_{exp} (focus of excitation in Explorer), which is not within the target area. Other neurons, most of which are located in between the G_1 and G_{exp}, are colored in gray (darker gray neurons are closer to G_{exp}). **Bottom left:** activities of Explorer neurons. **Rightmost three graphs on top:** activities of neurons from D1, SNr, and M before learning (i.e., before the target was hit). Rightmost three graphs on bottom: activities of the same neurons after learning, i.e., after the target was hit several times. Note that the maximum of the vertical axis of SNr is 0.5, while that of the other graphs is one. Horizontal dashed line of graphs for M (right) represents η.

and Thalamus, causing more excitation to neuron s_G in D1 and D2, even further disinhibition of the feedback loop, and further inhibition of the loops of other neurons. BG-mediated bias was in favor of movements toward G_1, implemented by an increase in weights from Context and Cortex to the neurons in D1 and D2 that correspond to a movement to G_1 (Equation 1). Thus, Cortex neuron s_G increased above η and movement was made to the location corresponding to G_1, even though the Explorer more-strongly excited neurons corresponding to movements made to G_{exp}.

Movement redistribution under contextual bias

The biasing of activity within the BG, BG's regulation of Cortex-Thalamus loop excitability, and the gradual focusing of excitation from Explorer to Cortex, comprise simple mechanisms that results in a seemingly "intelligent" structured transition from variability to repetition. After the target had been hit by chance a few times, weights from Context to neurons s_G in D1 and D2, and weights from neuron s_G in Cortex to neurons s_G in D1 and D2, were increased a little (Equation 1). When Explorer later chooses G_{exp} near G_1, the resulting relatively high excitation to Cortex neuron s_G, combined with the increased gain at Cortex-Thalamus loop s_G and decreased gain to other loops, excited Cortex neuron s_G while preventing other Cortex neurons from increasing past η. Thus, a movement to the target was made when Explorer chose G_{exp} near G_1: the target was hit with an increased likelihood, and movements to areas near the target were made with a decreased likelihood. We refer to this pattern as a "bias zone," centered at G_1, that increases in size the more often the target is hit.

Figure 4 shows how the bias zone increases as the number of times the target has been hit increases. In order to produce this figure, the model was run with G_{exp} set to G_1 for a set number of times. Then, learning was turned off and model response for G_{exp} set to each possible location was examined. Each graph in **Figure 4** plots the location of G_{exp} in the workspace: green dots indicate locations of G_{exp} that result in movements made to those locations; red dots indicate locations of G_{exp} that result in movements made to locations within the target area (red circle). The title of each graph indicates how many times G_{exp} was set to G_1 before response was examined. The expansion of the bias zone determines an "intelligent-looking" structured transition from variation to repetition in that it follows a non-random pattern.

For the purposes of this paper, model behavior is considered to be well-learned when a "streak" of hitting the target with ten consecutive movements is achieved. **Figure 5**, top left, plots the proportion of 50 runs that achieved this streak by various points of experience. About 40% reached it by 100 movements, and almost 80% reached it by 300 movements. A little over 20% did not achieve it by 300 movements. **Figure 5**, bottom left, plots the proportion of 50 runs that hit the target as a function of movement number. The proportion reaches about 0.8 by movement number 300.

Figure 5, right, plots, for each movement across the 50 runs, the distance between the movement and G_1 as a function of movement number. The distance of movements that hit G_1 are plotted in red (and are all at zero). As movement number increases, the density of movements near G_1 but that did not hit G_1 decreases at a faster rate than the density of movements far from G_1. This pattern is due to the expanding bias zone (**Figure 4**). We develop

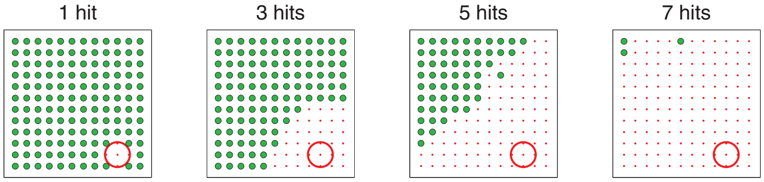

FIGURE 4 | Illustration of the "bias zone" effect. In each graph, the target was first hit *N* times (labeled at the top of the graph). Then, learning was turned off and movement for each possible value of G_{exp} was evaluated. Each dot represents the spatial location corresponding to G_{exp}. Large green dots represent locations of G_{exp} that resulted in movements that hit the location corresponding to G_{exp}. Small red dots represent locations of G_{exp} that, because of biasing implemented by weights onto D1 and D2, resulted in movements that hit the target (represented by the red circle in the lower right).

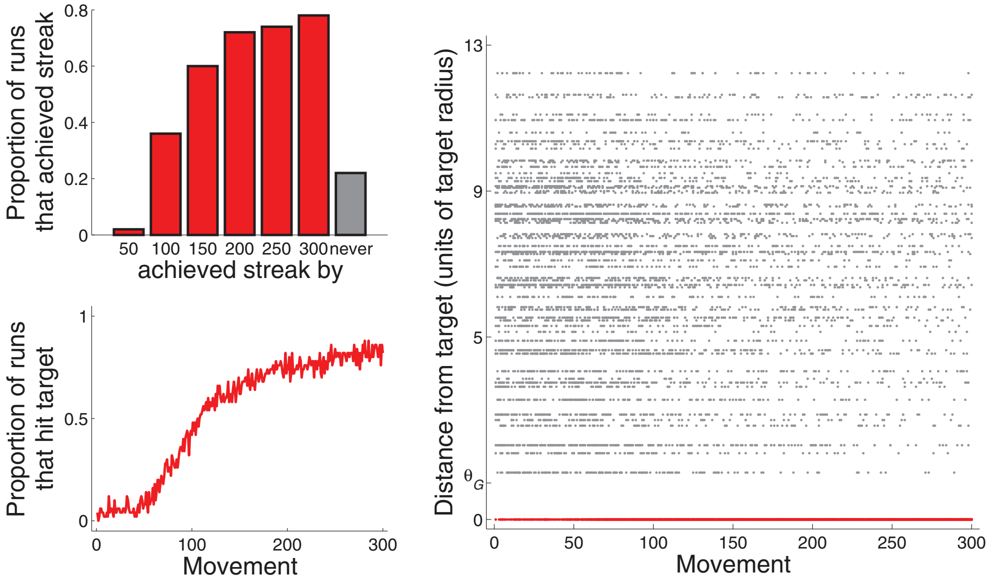

FIGURE 5 | Performance across all 50 runs for Experiment 1: A single target (and only BG biasing). Top left: proportion of runs that achieved streak of hitting target ten consecutive times by the movement 50, 100, 150, ..., or 300. Note that the bar graphs are cumulative. **Bottom left:** proportion of runs that hit target as a function of movement number. **Right:** Distance from the center of the target (in units of target radius) of each movement from all 50 runs. That for movements that hit the target are drawn in red and are at value 0 of the vertical axis.

a method for quantifying this pattern in the section describing results of Experiment 5 (and in the Supplementary section). Experiment 4 describes behavior in a more complicated task that results from this pattern.

Effect of cortical noise on model performance

The capability of the model to bias movements toward G_1 is due in part to the pattern of excitation from Explorer to Cortex (**Figure 2**), which weakly-excites all Cortex neurons by very similar amounts early in a trial. This suggests that model performance may be sensitive to unpredicted deviations from this pattern. To investigate this, we ran simulations in which signal-dependent noise (Harris and Wolpert, 1998) was added to Cortex neurons (which project to the BG and Thalamus, and from which movement is determined). In particular, at each time step: $y \leftarrow [y + y\,N(0,\ \sigma)]_0^1$, where y is the output activity of a Cortex neuron, $N(0,\ \sigma)$ refers to a number drawn randomly from a zero-mean Gaussian distribution with standard deviation σ, and $[x]_0^1$ returns 0 if $x < 0$, 1 if $x > 1$, and x otherwise. The proportion of the last 30 movements of all runs under a particular noise condition that were made to G_1 were 0.82, 0.64, 0.53, and 0.20 for σ levels of 0 (no noise), 0.1, 0.3, and 0.5, respectively. Thus, the model was able to learn to repeatedly hit G_1 if a low to moderate level of noise was added to Cortex neuron activity, but performance dropped off with high levels of noise. **Figure 6** illustrates, in a manner similar to **Figure 3**, example model neuron activity for a model run with $\sigma = 0.1$. The rest of the simulations in this paper were run with no noise.

3.2. EXPERIMENT 2: TWO SIMULTANEOUS TARGETS (G_1 AND G_{2FAR})

Movements that hit either of two targets, G_1 (lower right of the workspace) or G_{2far} (upper left) (red and blue circles, respectively, in **Figure 7**), were reinforced according to Equation 1. However, the habituation term differentiated them. (The habituation term is β^{N_k-1} in Equation 1, where N_k is the number of times target k has been hit and $\beta = 0.825$.) For example, even if G_1 was hit many times, at the first time G_{2far} was hit, it was a novel event and thus the corresponding weights increased by a large amount.

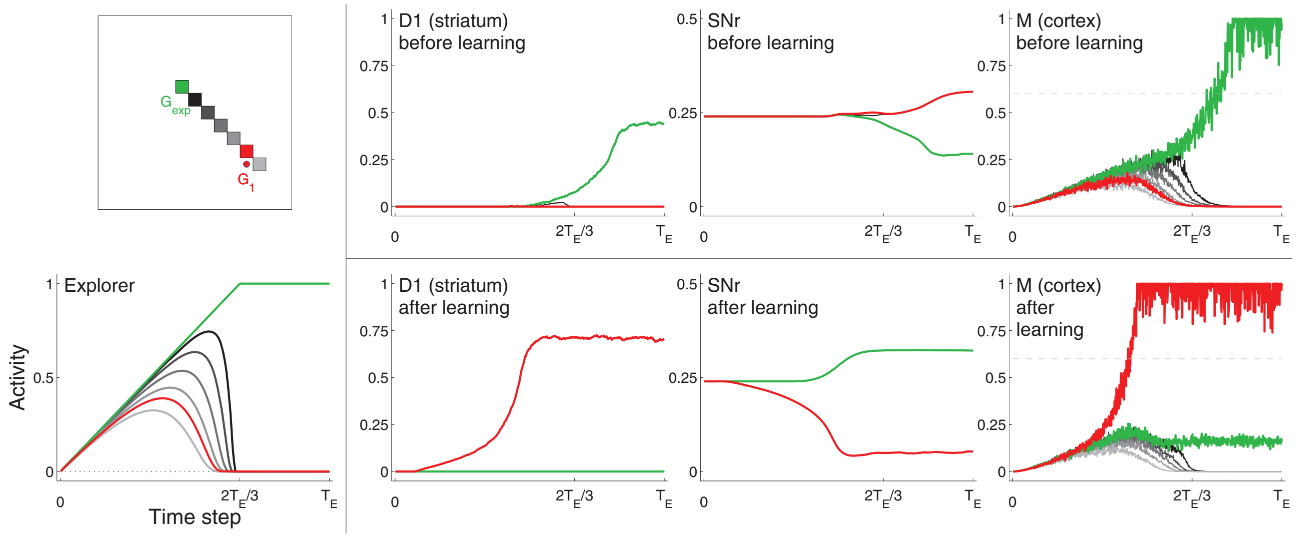

FIGURE 6 | Example neural activity for selected neurons from D1, SNr, and Cortex (M) at different points in training for a model with low levels of noise ($\sigma = 0.1$) added to Cortex (M) neuron activity (see text for details). This figure is plotted in a manner similar to that of **Figure 3**. Neurons are colored according to their spatial location in the grid (top left). The red neuron corresponds to the reinforced movement that hit target G_1 in this example. Note, however, that, unlike with **Figure 3**, the reinforced movement is one unit away from the center of G_1 (the center of G_1 is marked with a closed red circle). (Recall that the radius of the target is 1.1 units, so movements made to the center of G_1 or to the immediate neighbors of the center are reinforced.) The green neuron corresponds to the location of G_{exp} (focus of excitation in Explorer), which is not within the target area.

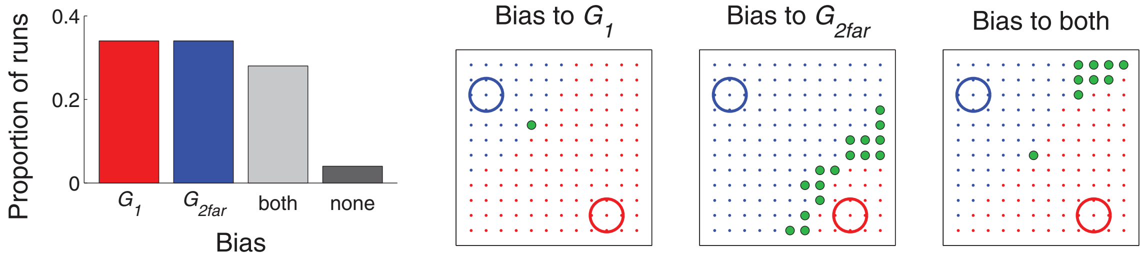

FIGURE 7 | Distribution of movements in Experiment 2: two simultaneous targets. Left: Proportion of runs that were classified as being biased toward one target (more than half of the movements hit target G_1, red, or G_{2far}, blue); both targets (more than a quarter of the movements hit G_1 and more than a quarter hit G_{2far}, gray); or none (all others, black). **Right three graphs:** example movement distributions for runs that were classified as being biased toward G_1, G_{2far}, or both. Similar to **Figure 4**, learning was stopped (this time at the end of the 300 movements) and then movement resulting from each possible value of G_{exp} (represented by the spatial locations of the dots) was evaluated. Red dots indicate the location of G_{exp} that resulted in a movement made to G_1 (red circle); blue dots indicate a movement made to G_{2far} (blue circle), and green dots indicate a movement made to G_{exp}.

Figure 7, left, plots the proportion of runs that were classified as either biased toward one of the targets, distributed between the two targets, or did not find a target (see figure caption for details on the classification criteria). While behavior in a majority of the runs was biased to a single target (e.g., middle two graphs of **Figure 7**), the model was capable of distributing movements to both targets (e.g., **Figure 7**, right). For runs which were biased to just one target, only a G_{exp} very near the un–preferred target produced a movement to that target.

3.3. EXPERIMENT 3: REINFORCE G_1, THEN G_{2FAR}, THEN G_1 AGAIN

The use of experience-based learning rules—weight modification (Equation 1) is dependent on actual behavior—and a habituation term leads to a type of memory that can influence subsequent behavior in a changing environment. This is illustrated with experiments in which only movements to G_1 are reinforced for 300 movements, then only movements to G_{2far} are reinforced (at which point the habituation term for G_1 is reset), and then only movements to G_1 are reinforced again. As shown in **Figure 8**, top row, which plots the proportion of runs that hit each target as a function of movement number, the reacquisition of G_1 (movements 601–900) occurred faster than the initial acquisition (movements 1–300) of G_1.

The enhanced acquisition is because corticostriatal weights corresponding to movements toward G_1, illustrated in red in **Figure 8**, bottom row, increased to a stable value (of about 0.2 in the figure) during first acquisition. (The habituation prevents it from increasing any more after the target had been

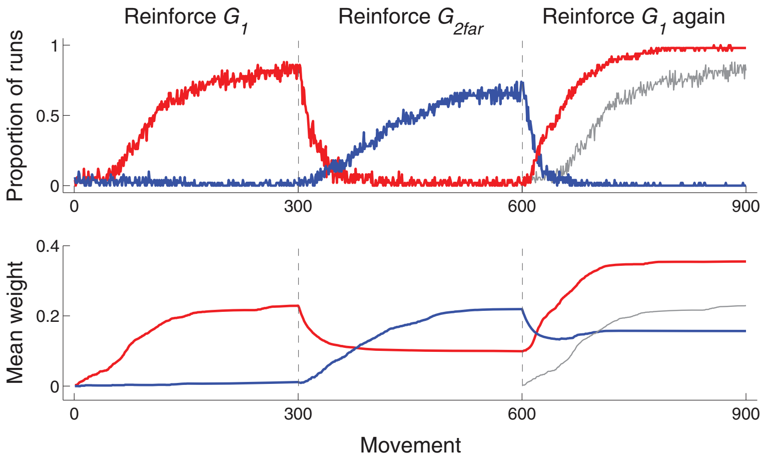

FIGURE 8 | Time-course of behavior and corticostriatal weights for Experiment 3: reinforce G_1 (movements 1–300), then G_{2far} (301–600), then G_1 again (601–900). Top: proportion of runs that hit G_1 (red) or G_{2far} (blue) as a function of movement number. The proportion of runs that hit G_1 during movements 1–300 are redrawn at horizontal positions 601–900 as a gray line for comparison of performance between initial acquisition (movements 1–300) and reacquisition (movements 601–900) of G_1. **Bottom:** Mean (across runs) weight from Context neuron to the D1 neuron corresponding to most movements that hit G_1 (red) or G_{2far} (blue) for that particular run. The D1 neuron that corresponded to most movements that hit each target was determined by finding the maximum weight from Context to D1 neurons at the end of each 300 movement segment. Because several movements can hit each target, only runs in which the same D1 neuron was selected at movement 300 and movement 900 (i.e., for movements that hit G_1) were included (16 out of 50 runs were excluded). That for weights from Context neuron to D2 neurons followed a similar pattern and are not plotted. Similar to the graphs in the top row, mean weight during movements 1–300 are plotted again at movements 601–900 in gray for comparison purposes.

repeatedly hit.) During movements 301–600, G_{2far} was reinforced (and G_1 was no longer reinforced). The model continued to move to G_1 early in the second set of movements, but, because G_1 was no longer reinforced, the corresponding weights decreased. As the weights decreased, the bias zone around G_1 decreased and the model was free to move to other locations, including toward G_{2far}. As a new bias zone, now centered on G_{2far}, was established, the model stopped moving to G_1. Because movements toward G_1 were no longer made, weights associated with moving to G_1 ceased to decrease. When movements to G_1 were reinforced again, those weights were already above zero and thus G_1 was reacquired faster than it was initially acquired. In addition, due to resetting the habituation term, the weights increased to a greater value than the previous high value.

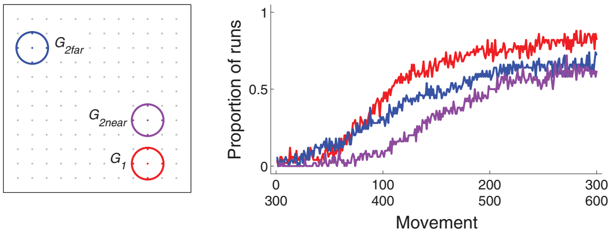

FIGURE 9 | Behavior for Experiment 4: reinforce G_1 (movements 1–300) and then either G_{2far} or G_{2near} (301–600). Left: locations of the three targets in the workspace (dots indicate locations corresponding to possible values of G_{exp}, colored gray if those locations do not lie within a target area. **Right:** proportion of runs that hit G_1 for movements 1–300 or G_{2far} or G_{2near} for movements 301–600.

This pattern of activity provides a simple mechanism that can be used to partially explain the findings that practice sessions that are separated in time lead to enhanced acquisition and performance compared to practice sessions that are massed together (Ammons, 1950; Baddeley and Longman, 1978) (though such effects do not necessarily apply to all types of tasks, e.g., Lee and Genovese 1989).

3.4. EXPERIMENT 4: REINFORCE G_1, THEN EITHER G_{2FAR} OR G_{2NEAR}

When one target is reinforced for a period of time, and then another is reinforced instead, how well the second reinforced target is acquired depends on its proximity to the first target. This is illustrated by comparing the results of experiments in which the second target (G_{2far}, blue in **Figure 9**) was far from the first one with those in which the second target (G_{2near}, purple) was near the first one. **Figure 9** plots the proportion of runs for which the first and second targets were hit as a function of movement for the different sets. The first target (G_1, red) was acquired the fastest. The far second target (G_{2far}) was acquired faster than the near second target (G_{2near}).

The discrepancy between acquiring the second targets is explained by the bias zone. A well-learned model has corticostriatal weights such that the bias zone is large. When the bias zone is centered around G_1, un–reinforced movements to G_1 must happen in order for weights to decrease, after which the bias zone shrinks and movements to other locations can be made. Movements to locations far from G_1 are available earlier than movements to locations near G_1 as the bias zone shrinks. Thus, a second target far from G_1 will be more-easily acquired than a second target near G_1.

3.5. EXPERIMENT 5: MOVEMENT REDISTRIBUTION UNDER DIFFERENT BIAS CONDITIONS

As movements made to a target increase, movements made to other locations must decrease: movements are redistributed over the workspace. The previous sections focused on movement redistribution in our model with only BG-mediated biasing (Equation 1). Here we describe metrics of movement redistribution that will allow us to compare how movements are redistributed under different bias conditions. We focus on model runs in which only movements made to one target (G_1) were reinforced.

Redistribution metric

The expanding bias zone (Results section 3.1 and **Figure 4**) that results from BG-mediated biasing results in a pattern of behavior such that movements made near, but not at, the target decrease in likelihood earlier than movements made far from the target. For each run, we quantify the rate of decrease as a function of distance from target. Briefly (see **Figure 10**), movements that did not hit the target were coarsely categorized into three temporal chunks and three spatial zones (vertical and horizontal lines, respectively, in **Figure 10**). Temporal chunk one includes the first 100 movements; temporal chunk two includes the second 100 movements; and temporal chunk three includes the last 100 movements. Recalling that θ_G is target radius and letting dX be the distance of a movement from target center, the spatial zones are 1) $\theta_G < dX \leq 5\theta_G$ (green points in **Figure 10**), 2) $5\theta_G < dX \leq 9\theta_G$ (blue), and 3) $9\theta_G < dX$ (black). The number of movements that fell into spatial zone i from temporal chunks 1 to 2 to 3 was fit to an equation of the form $e^{b_i(j-1)}$, where j refers to temporal chunk. The rate of decrease of the number movements was quantified by the parameter b_i. A more negative b_i indicates a greater rate of decrease (see the Supplementary section for more details).

Movement redistribution across different bias conditions

Figure 10, top row, graphs movement distance from target as a function of movement number for three sample runs under BG-mediated bias (these graphs are similar to **Figure 5**, right). In all three cases, $b_1 < b_2 < b_3$, i.e., the rate of decrease of movements made near but not at the target is greater than that of movements made far from the target. This is in line with the behavioral pattern we would expect given the expanding bias zone (**Figure 4**) that results from BG-mediated biasing. Regarding the specific sample runs in **Figure 10**, top row, the rate of decrease of movements from the first sample run that fell within zone one is greater than that of the second sample run, which is greater than that of the third sample run. This, also, is reflected in the b metrics.

The same process was used to determine b metrics for models that biased movement selection with different mechanisms. Recall from Methods section 2.5 that, if there is no "Cognitive bias," movements suggested by the Explorer (G_{exp}) were randomly selected from a uniform distribution over all possible movements. Under the Cognitive bias scheme (described in Methods section 2.5 and the Supplementary section), every time the target is hit, the set of possible movements from which G_{exp} is selected decreases: movements further from target center are removed from the set earlier than movements closer to target center. Movement redistribution under a Cognitive bias thus follows a trend opposite that under BG-mediated bias: $b_1 > b_2 > b_3$ (**Figure 10**, bottom row).

For a given run of a model using BG-mediated bias, b for spatial zones closer to the target should be more negative than b for zones farther from the target. Thus, we expect $b_3 - b_2 > 0$ and $b_2 - b_1 > 0$ in models using BG-mediated bias. Models using the Cognitive bias should exhibit opposite behavior: $b_3 - b_2 < 0$ and $b_2 - b_1 < 0$. The differences should be zero if the transition from variation to repetition does not follow a structured pattern (i.e., the frequency of movements to non-target areas decreases uniformly).

Figure 11 plots the distribution of pair-wise (by run) differences $b_3 - b_2$ (right column, black) and $b_2 - b_1$ (left column, blue) of model runs using different bias conditions (arranged by row). The means of the distributions were also tested against the null hypothesis that they are zero (single sample one-tailed t-tests). The distributions of the pair-wise differences for models using a BG bias (top row) were positive; that for models using a Cognitive bias (bottom) were negative; and that for using both biasing mechanisms (middle) were also negative (though visual inspection suggests that the Cognitive bias condition has more extreme negative pair-wise differences than does the combined bias condition). Thus, this analysis was able to capture the general trends that were seen in the different bias conditions of the model.

4. DISCUSSION

As described in a recent theory of *action discovery* (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013), when an unexpected sensory event occurs, animals transition from executing a variety of movements to repeating movements that may have caused the event. A transition from variation to repetition often follows non-random, structured patterns that may be explained with sophisticated cognitive mechanisms (e.g., Dearden et al. 1998; Dimitrakakis 2006; Simsek and Barto 2006). However, in action discovery, simple non-cognitive mechanisms involving dopamine modulation of basal ganglia (BG) activity are thought to play a prominent role in behavioral biasing. In this paper we use a biologically-plausible computational model to demonstrate that a structured transition from variation to repetition can emerge from processing within such simple mechanisms. Such behavior is due to the following features on which our model, unlike most previous models of BG function, focuses: (i) the BG does not bias behavior directly, but modulates cortical response to excitation (Chevalier and Deniau, 1990; Mink, 1996; Humphries and Gurney, 2002; Cohen and Frank, 2009; Redgrave et al., 2011; Baldassarre et al., 2013); (ii) excitation to cortex follows a pattern that evolves from weakly exciting all neurons to strongly exciting only one neuron (Britten et al., 1992; Platt and Glimcher, 1999; Huk and Shadlen, 2005; Bogacz et al., 2006; Gold and Shadlen, 2007; Lepora et al., 2012). By including these features in our model, we show that sophisticated cognitive mechanisms may not always be necessary to develop a structured transition from variation to repetition.

In our model, movements occur by selecting an end-point (spatial location) to which to move. Movements that terminated

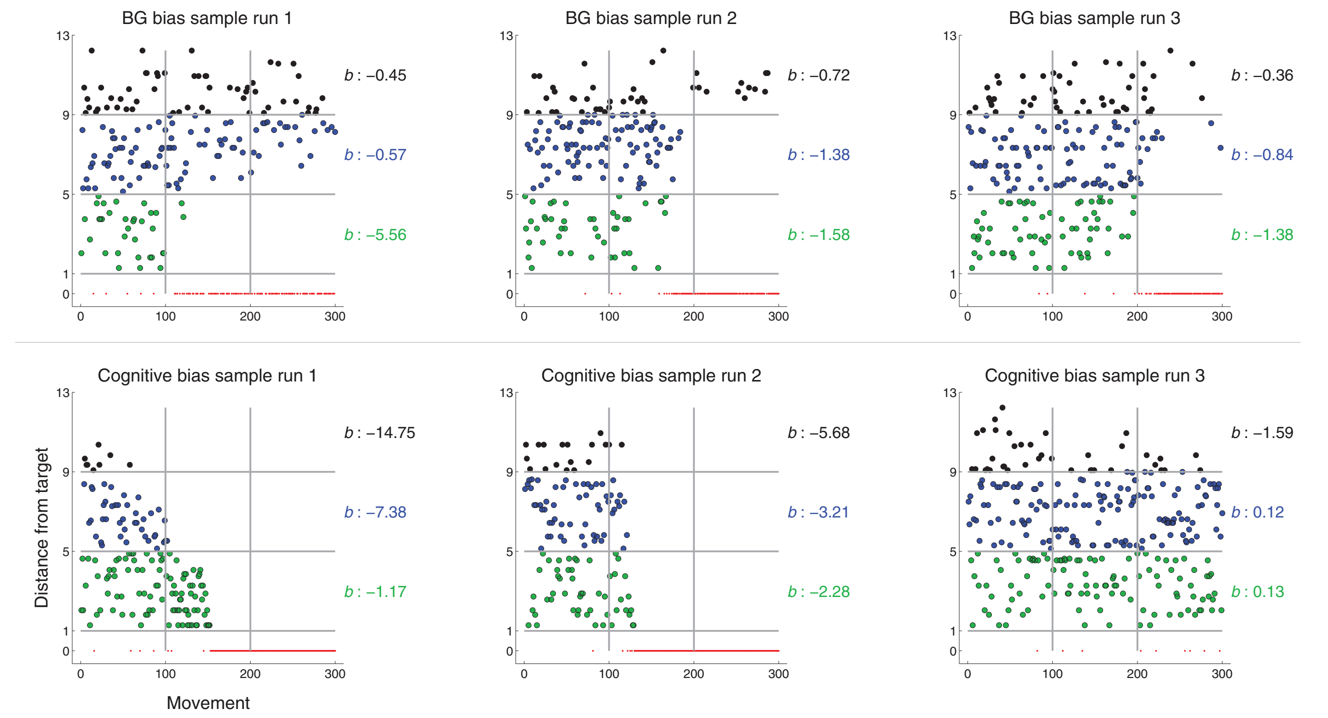

FIGURE 10 | Distance from G_1 of executed movement (in units of θ_G) as a function of movement number (similar to Figure 5, right) for sample runs of models using BG-mediated biasing (top row) or Cognitive biasing (bottom). Color of dot indicates the spatial zone (defined relative to target center) in which the movement lies. Horizontal lines indicate spatial zone borders. Vertical lines indicate temporal chunk borders. The parameter *b* indicates the rate of decrease of movements falling within each spatial zone. The more negative *b* is, the greater the rate of decrease (see text for details).

in a target area were reinforced so that the selection of such end-points increased in frequency. The transition from executing a variety of movements to executing just the reinforced movements followed a structured pattern: as end-points at the target location increased in frequency, end-points near, but not at, the target location decreased in frequency at a greater rate than end-points far from the target. We refer to the area around the target area in which end-point frequency decreased as a "bias zone" (**Figures 4**, **10**, top), and the bias zone increased in size as the target was repeatedly hit. The graded shift from variation (a small bias zone) to repetition (a large bias zone) allows for the discovery of a second target area in some cases (**Figure 7**), and also results in specific patterns of behavior if the target area moves (**Figures 8**, **9**).

In addition, in action discovery, phasic DA activity in response to achievement of the outcome (e.g., hitting the reinforced target area) decreases as associative brain areas learn to predict the outcome's occurrence (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013; Mirolli et al., 2013). This may be thought of as a type of intrinsic motivation (IM) in that the outcome need not have hedonic value in order to be reinforcing (Oudeyer and Kaplan, 2007; Baldassarre, 2011; Barto, 2013; Barto et al., 2013; Gottlieb et al., 2013; Gurney et al., 2013). The type of IM in action discovery is best described as some combination of novelty and surprise (Barto et al., 2013). A detailed account of exactly how the prediction process may be implemented in the brain is beyond the scope of this paper. We mimic its effects in our model with a simple habituation mechanism similar that used in neural network models of novelty detection (Marsland, 2009). Here, the reinforcing effects of an outcome with which the model has little recent experience is greater than the reinforcing effects of an outcome with which the model has much recent experience. The habituation term ($\beta^{N_k - 1}$ in equation 1) influences behavioral patterns, particularly in tasks in which more than one target area is reinforced (**Figure 7**) or the target area changes (**Figures 8**, **9**). Unlike the reward prediction error hypothesis of phasic DA neuron activity (Houk et al., 1995; Schultz et al., 1997), habituation is a mechanism that does not rely on extrinsic motivation by which phasic DA neuron activity, and hence rate of change of the rate of corticostriatal plasticity, decreases with continued occurrences of the outcome.

We also implement models in which a structured transition from variation to repetition is that which would be expected if one type of more sophisticated mechanism ("Cognitive biasing") is in effect. The pattern of behavior (**Figure 10**, bottom) is then different than that of BG-only biasing. Finally, we have devised a method for capturing such differences with quantitative measures (**Figures 10**, **11**) which will allow us to make contact with future behavioral experiments investigating how different brain areas contribute to biasing behavior in tasks similar to model tasks. In continuing work, we are devising such behavioral experiments. Preliminary results suggest that our quantitative measure will allow us to compare the effects of different biasing mechanisms

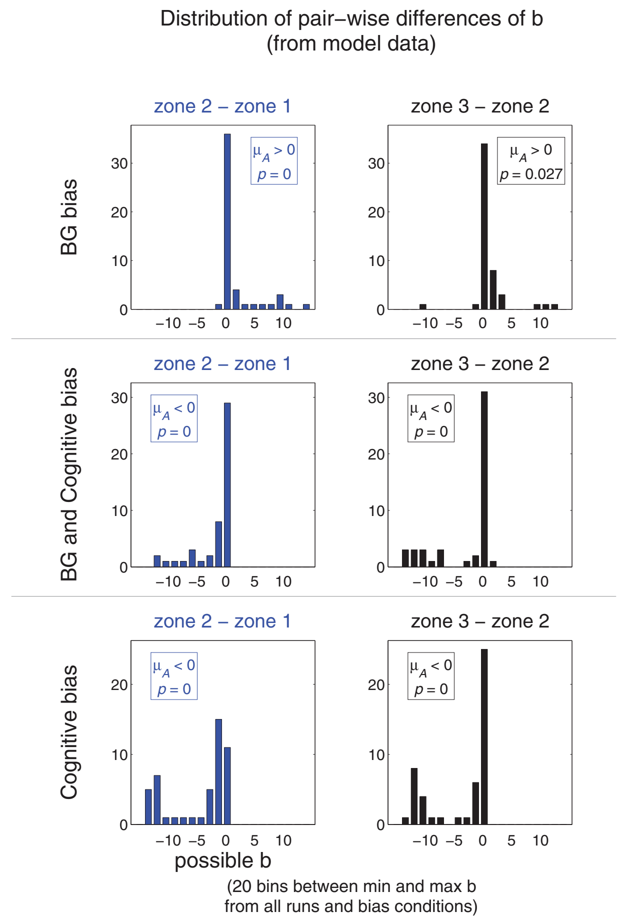

FIGURE 11 | Comparing redistribution patterns in Experiment 5: movement redistribution under different bias conditions. Each graph is a histogram illustrating the distribution of pair-wise differences (by run) of b parameters of model runs using different bias conditions (arranged by row). That for $b_3 - b_2$ is in the right column (black); that for $b_2 - b_1$ is in the left column (blue). Bin widths and locations were determined as follows: the minimum and maximum b of all possible b from all analyzed runs from all conditions defined the range of possible b. This range was divided into 20 evenly-spaced bins of uniform size. The means of the samples of b parameters were tested according to the hypotheses that they have a mean $\mu_A > 0$ or $\mu_A < 0$ (indicated, along with p value, in each graph).

by examining behavior from different systems (e.g., model versus human), different workspaces, different target sizes, and different target locations, etc. Possible mechanisms by which to isolate different brain mechanisms include explicit instructions, use of different stimuli (Thirkettle et al., 2013b), or use of distractor tasks (Stocco et al., 2009).

As with any computational model of brain systems, the mechanisms described in this paper should be viewed as being a part of a complex system of interacting parts. We've isolated the effects of the specific mechanisms we've investigated in order to demonstrate how a structured transition from variation to repetition can emerge from those mechanisms. In the next subsection we discuss the implications of some of these choices in greater detail and how to expand on them to include more sophisticated systems.

4.1. A MULTI-STAGE SELECTION PROCESS

Recall that, for each movement in our model, the pattern of excitation from "Explorer" to "Cortex" evolves from weakly-exciting all neurons to strongly-exciting one neuron (referred to as G_{exp}, the focus of excitation). The weak excitation of all neurons early in the evolution allows for corticostriatal plasticity to bias behavior. Behavior can also be biased by the choice of G_{exp}, the effects of which are greater later in the evolution. Thus, the evolving excitation pattern from Explorer to Cortex allows for a multi-stage selection process. We expand on these points below.

Through corticostriatal plasticity and BG selection mechanisms, Cortex neurons that are only weakly excited during the early stages of excitation from Explorer can increase in activity at a greater rate than other Cortex neurons. BG selection mechanisms also enable these neurons to suppress the responses of other Cortex neurons to subsequent strong excitation (e.g., **Figure 3**). The expanding bias zone (described in Results section 3.1 and **Figure 4**) that is seen in models using BG-mediated biasing emerges from the pattern of excitation from Explorer to Cortex. Because the model task was a spatial reaching task, a topographic representation was used that revealed an apparent dependency between movements: neurons in Explorer near the focus of excitation (G_{exp}) were excited more than neurons far from the focus.

However, a different pattern may be revealed in other types of tasks. In general, the pattern of activity is likely to be influenced by perceptual processing of sensory information. For example, the theory of affordances (Gibson, 1977, 1986) suggests that the perception of objects preferentially primes neurons that correspond to actions that can operate on those objects, e.g., the perception of a mug would prime a grasping action. Thus, the pattern of excitation in these conditions would preferentially excite those neurons, and excitation may follow a pattern that is different than the one used in this paper. Because BG modulates how Cortex responds to excitation rather than directly-exciting movements, any behavioral pattern controlled by BG-mediated biasing would depend on the pattern of excitation to Cortex. Thus, different patterns of exploration, and different patterns of a structured transition from variation to repetition, would be observed in different environments and tasks.

We envision that more sophisticated mechanisms (e.g., our Cognitive biasing) can be expressed in our model in the later part of the evolving excitation pattern of the Explorer, i.e., in how G_{exp} is chosen. One such mechanism may search the workspace in a way that is more intelligent than random, such as a spiral or raster-like search pattern that does not repeat itself until all possible movements have been executed. The choice of G_{exp} could also be adaptive, including using mechanisms by which a transition from variation to repetition is governed by mechanisms based on measures of optimality, uncertainty, or other task-related variables (Dearden et al., 1998; Daw et al., 2006; Dimitrakakis, 2006; Simsek and Barto, 2006; Cohen et al., 2007).

Thus, the early part of the evolving excitation pattern from Explorer to Cortex comprises weak excitation that is influenced by perception of the environment (e.g., affordances or, in our model, possible movement locations) or simple mechanisms. The later part of the evolution allows for more complicated mechanisms that may require more processing time to also influence behavior. We have focused mostly on simple mechanisms in this paper, but the evolving pattern of excitation can be used to implement proposed theories that focus on multiple influences on behavior, e.g., Kawato (1990); Rosenstein and Barto (2004); Daw et al. (2005); Shah and Barto (2009).

4.2. ACTION DISCOVERY WITH COMPLICATED BEHAVIORS

There are many types of movements or behaviors that can affect the environment, e.g., making a gesture (regardless of spatial location), manipulating objects in the environment, or making a sequence of movements. In this paper we focused on a simple type of action in which the system, able to select a spatial end-point of movement, must discover the end-point(s) that delivers an outcome. On a more abstract level, this is similar to "*n*-armed bandit" problems, in which the system must discover which out of a set of *n* actions is followed by the most rewarding consequences in a one-step decision task (e.g., Sutton and Barto 1998). The general process of action discovery (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013) is also concerned with discovering the temporal and structural components of a complex behavior that affects the environment. These problems are similar to the those of temporal and structural credit assignment problems (Minksy, 1961; Sutton, 1984, 1988; Barto, 1985; Sutton and Barto, 1998), which we briefly describe below.

One form of the temporal credit assignment problem is exposed in systems in which a series of actions is required in order to achieve an outcome, and there is great redundancy: a large number of different (but possibly overlapping) sequences can achieve the outcome. How does the agent discover the most direct sequence, i.e., the sequence that uses the fewest actions? This redundancy is often resolved by assigning a cost for each executed action and using optimal control methods to achieve the goal while also minimizing cost (e.g., Sutton and Barto 1998). However, optimal control methods, which are designed to find behavior that minimizes cost according to an arbitrary cost function, may use mechanisms that are more sophisticated and complicated than those thought to underly action discovery. Recent modeling work (Shah and Gurney, 2011; Chersi et al., 2013) has shown that a simpler learning rule that does not incorporate cost per action can discover the most direct sequence of actions in a redundant system. Such behavior remains stable for a period of time, but, if learning is not attenuated, extraneous actions are incorporated with extended experience (Shah and Gurney, 2011).

The structural credit assignment problem is exposed when a system can execute many actions simultaneously and the outcome depends only on the simultaneous execution of a small subset of those. When behavior is composed of several components, and the outcome is contingent on only some of those components, variation allows the animal to determine which components are relevant and to "weed out" the irrelevant components. We have not addressed this problem directly, but previous work on the structural credit assignment problem in RL offers promising directions (Barto and Sutton, 1981; Barto et al., 1981; Barto, 1985; Barto and Anandan, 1985; Gullapalli, 1990).

4.3. CONCLUSION

How biasing causes a transition from variation to repetition so as to converge on the specific movements that cause an outcome is a fundamental problem in the process of action discovery. With a simple model of a restricted aspect of action discovery, which includes neural processing features not included in most other models of BG function, we are able to describe the effects of different types of behavioral biasing. The results reported in this paper describe a first step in understanding the more processes at work in general action discovery.

ACKNOWLEDGMENTS

We are grateful for the efforts of Martin Thirkettle (Open University and University of Sheffield) and Jennifer Tidman (Open University) who, along with Ashvin Shah, developed and ran the preliminary behavioral experiments referred to in the Discussion. In addition, the authors had helpful discussions with Peter Redgrave, Tom Stafford, Jen Lewis, and Nicolas Vautrelle (all at the University of Sheffield), and Patricia Shaw (Aberystwyth University). Finally, we thank anonymous reviewers for their careful reading and comments.

FUNDING

We are grateful for financial support from the European Union's Seventh Framework Programme grant FP7-ICT-IP-231722 (IM-CLeVeR: Intrinsically Motivated Cumulative Learning Versatile Robots).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00091/abstract

REFERENCES

Ammons, R. (1950). Acquisition of motor skill: III. effect of initially distributed practice on rotary pursuit performance. *J. Exp. Psychol.* 40, 777–787. doi: 10.1037/h0061049

Anderson, R., and Buneo, C. (2002). Intentional maps in posterior parietal cortex. *Ann. Rev. Neurosci.* 25, 189–220. doi: 10.1146/annurev.neuro.25.112701.142922

Baddeley, A., and Longman, D. (1978). The influence of length and frequence of training session on the rate of learning to type. *Ergonomics* 21, 627–635. doi: 10.1080/00140137808931764

Baldassarre, G. (2011). "What are intrinsic motivations? A biological perspective," in *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, eds A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P. Y. Oudeyer, et al. (Piscataway, NJ: IEEE), E1–E8.

Baldassarre, G., Mannella, F., Fiore, V., Redgrave, P., Gurney, K., and Mirolli, M. (2013). Intrinsically motivated action-outcome learning and goal-based action recall: a system-level bio-constrained computational model. *Neural Netw.* 41, 168–187. doi: 10.1016/j.neunet.2012.09.015

Barto, A. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Hum. Neurobiol.* 4, 229–256.

Barto, A. (2013). "Intrinsic motivation and reinforcement learning," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 1, eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer-Verlag), 17–47. doi: 10.1007/978-3-642-32375-1-2

Barto, A., and Anandan, P. (1985). Pattern-recognizing stochastic learning automata. *IEEE Trans. Syst. Man Cybern.* 15, 360–375. doi: 10.1109/TSMC.1985.6313371

Barto, A., Mirolli, M., and Baldasarre, G. (2013). Novelty or surpise? *Front. Psychol.* 4:1. doi: 10.3389/fpsyg.2013.00907

Barto, A., and Sutton, R. (1981). Landmark learning: an illustration of associative search. *Biol. Cybern.* 42, 1–8. doi: 10.1007/BF00335152

Barto, A., Sutton, R., and Brouwer, P. (1981). Associative search network: a reinforcement learning associative memory. *Biol. Cybern.* 40, 201–211. doi: 10.1007/BF00453370

Bertsekas, D., and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. (2006). The physics of optimal decision making: a formal analysis of models of performance in twi-alternative forced-choice tasks. *Psychol. Rev.* 113, 700–765. doi: 10.1037/0033-295X.113.4.700

Bolam, J., Bergman, H., Graybiel, A., Kimura, M., Plenz, D., Seung, H., et al. (2006). "Group report: microcircuits, molecules, and motivated behavior—microcircuits in the striatum," in *Microcircuits: The Interface Between Neurons and Global Brain Function*, Dahlem Workshops Reports, Chapter 9, eds S. Grillner and A. Graybiel (Cambridge, MA: MIT Press), 165–190.

Britten, K., Shadlen, M., Newsome, W., and Movshon, J. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurophysiol.* 12, 4745–4765.

Calabresi, P., Picconi, B., Tozzi, A., and DiFilippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci.* 30, 211–219. doi: 10.1016/j.tins.2007.03.001

Chambers, J., Gurney, K., Humphries, M., and Prescott, T. (2011). "Mechanisms of choice in the primate brain: a quick look at positive feedback," in *Modelling Natural Action Selection*, Chapter 17, eds A. Seth, T. Prescott, and J. Bryson (Cambridge: Cambridge University Press), 390–418.

Chersi, F., Mirolli, M., Pezzulo, G., and Baldessarre, G. (2013). A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning. *Neural Netw.* 41, 212–224. doi: 10.1016/j.neunet.2012.11.009

Chevalier, G., and Deniau, J. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci.* 13, 277–281. doi: 10.1016/0166-2236(90)90109-N

Cohen, J., McClure, S., and Yuo, A. (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 Biol. Sci.* 362, 933–942. doi: 10.1098/rstb.2007.2098

Cohen, M. X., and Frank, M. J. (2009). Neurocomputational models of the basal ganglia in learning, memory, and choice. *Behav. Brain Res.* 199, 141–156. doi: 10.1016/j.bbr.2008.09.029

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711. doi: 10.1038/nn1560

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

Dearden, R., Friedman, N., and Russell, S. (1998). "Bayesian q-learning," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)* (Madison, WI), 761–768.

Dimitrakakis, C. (2006). "Nearly optimal exploration-exploitation decision thresholds," in *Proceedings of the Sixteenth International Conference on Artificial Neural Networks (ICANN 2006), Part I, Athens, Greece*, eds S. Kollias, A. Stafylopatis, W. Duch, and E. Oja (Berlin; Heidelberg: Springer), 850–859.

Georgopoulos, A., Kalaska, J., Caminiti, R., and Massey, J. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.* 2, 1527–1537.

Gibson, J. (1977). "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, eds R. Shaw and J. Bransford (Hillsdale, NJ: Lawrence Erlbaum and Associates), 67–82.

Gibson, J. (1986). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Ann. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Gottlieb, J., Lopes, P. O. M., and Baranes, A. (2013). Information-seeking, curiousity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593. doi: 10.1016/j.tics.2013.09.001

Gullapalli, V. (1990). A stochastic reinforcement learning algorithm fo learning real-valued functions. *Neural Netw.* 3, 671–692. doi: 10.1016/0893-6080(90)90056-Q

Gurney, K., Lepora, N., Shah, A., Koene, A., and Redgrave, P. (2013). "Action discovery and intrinsic motivation: a biologically constrained formalism," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 7, eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer-Verlag), 151–184. doi: 10.1007/978-3-642-32375-1-7

Gurney, K., Prescott, T., and Redgrave, R. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol. Cybern.* 84, 401–410. doi: 10.1007/PL00007984

Gurney, K., Redgrave, R., and Prescott, T. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biol. Cybern.* 84, 411–423. doi: 10.1007/PL00007985

Gurney, K., Prescott, T. J., Wickens, J. R., and Redgrave, P. (2004). Computational models of the basal ganglia: from robots to membranes. *Trends Neurosci.* 27, 453–459. doi: 10.1016/j.tins.2004.06.003

Harris, C., and Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature* 394, 780–784. doi: 10.1038/29528

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). "A model of how the basal ganglia generate and use neural signals that predict reinforcement," in *Models of Information Processing in the Basal Ganglia*, Chapter 13, eds J. C. Houk, J. L. Davis, and D. G. Beiser (Cambridge, MA: MIT Press), 249–270.

Huk, A., and Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.* 25, 10420–10436. doi: 10.1523/JNEUROSCI.4684-04.2005

Humphries, M., and Gurney, K. (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network* 13, 131–156. doi: 10.1088/0954-898X/13/1/305

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* 15, 535–547. doi: 10.1016/S0893-6080(02)00047-3

Kawato, M. (1990). "Feedback-error-learning neural network for supervised motor learning," in *Advanced Neural Computers*, ed R. Eckmiller (North-Holland: Elsevier), 365–372.

Koos, T., and Tepper, J. (1999). Inhibitory control of neostriatal projection neurons by GABAergic interneurons. *Nat. Neurosci.* 2, 467–472. doi: 10.1038/8138

Lee, T., and Genovese, E. (1989). Distribution of practice in motor skill acquisition: different effets for discrete and continuous tasks. *Res. Q. Exerc. Sport* 60, 59–65. doi: 10.1080/02701367.1989.10607414

Lepora, N., Fox, C., Evans, M., Diamond, M., Gurney, K., and Prescott, T. (2012). Optimal decision-making in mammals: insights from a robot study of rodent texture discrimination. *J. R. Soc. Interface* 9, 1517–1528. doi: 10.1098/rsif.2011.0750

Marsland, S. (2009). Using habituation in machine learning. *Neurobiol. Learn. Mem.* 92, 260–266. doi: 10.1016/j.nlm.2008.05.014

Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* 50, 381–425. doi: 10.1016/S0301-0082(96)00042-1

Minksy, M. (1961). "Steps toward artificial intelligence," in *Proceedings IRE 49 1*, 8–30. (Reprinted in Computers and Thought, McGraw-Hill, 1963. Mapped out future research in AI with emphasis on symbolic descriptions). doi: 10.1109/JRPROC.1961.287775

Mirolli, M., Baldassarre, G., and Santucci, V. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcement driving both action acquisition and reward maximization: a simulated robotic study. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012

Oudeyer, P., and Kaplan, F. (2007). What is intrinsic motivation? A topology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007

Platt, M., and Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature* 400, 233–238. doi: 10.1038/22268

Prescott, T., Gonzalez, F., Gurney, K., Humphries, M., and Redgrave, R. (2006). A robot model of the basal ganglia: behavior and intrinsic processing. *Neural Netw.* 19, 31–61. doi: 10.1016/j.neunet.2005.06.049

Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339. doi: 10.1016/j.brainresrev.2007.10.007

Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., and Lewis, J. (2013). "The role of the basal ganglia in discovering novel actions," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 6, eds G. Baldassarre and M.

Mirolli (Berlin; Heidelberg: Springer-Verlag), 129–150. doi: 10.1007/978-3-642-32375-1-6

Redgrave, P., Vautrelle, N., and Reynolds, J. (2011). Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neurscience* 198, 138–151. doi: 10.1016/j.neuroscience.2011.07.060

Rosenstein, M., and Barto, A. (2004). "Supervised actor-critic reinforcement learning." in *Handbook of Learning and Approximate Dynamic Programming, IEEE Press Series on Computational Intelligence,* Chapter 14, eds J. Si, A. Barto, W. Powell, and D. Wunsch (Piscataway, NJ: Wiley-IEEE Press), 359–380.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Shah, A., and Barto, A. G. (2009). Effect on movement selection of an evolving sensory representation: a multiple controller model of skill acquisition. *Brain Res.* 1299, 55–73. doi: 10.1016/j.brainres.2009.07.006

Shah, A., and Gurney, K. (2011). "Dopamine-mediated action discovery promotes optimal behaviour 'for free,"' in *Twentieth Annual Computational Neuroscience Meeting*, Stockholm. (Poster presentation. Absract also published in BMC Neuroscience 2011, 12 (Suppl. 1):P138).

Simsek, O., and Barto, A. (2006). "An intrinsic reward mechanism for efficient exploration," in *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML-06)*, eds W. Cohen and A. Moore (Pittsburgh, PA: ACM Inernational Conference Proceedings Series), 833–840.

Skinner, B. (1938). *The Behavior of Organisms*. New York, NY: Appleton-Century-Crofts.

Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., et al. (2012). A novel task for the investigation of action acquisition. *PLoS ONE* 7:e37749. doi: 10.1371/journal.pone.0037749

Stafford, T., Walton, T., Hetherington, L., Thirkettle, M., Gurney, K., and Redgrave, P. (2013). "A novel behavioural task for researching intrinsic motivations," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 15, eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer-Verlag), 395–410.

Stocco, A., Fum, D., and Napoli, A. (2009). Dissociable processes underlying decisions in the Iowa Gambling Task: a new integrative framework. *Behav. Brain Funct.* 5, 1. doi: 10.1186/1744-9081-5-1

Sutton, R. (1984). *Temporal Credit Assignment in Reinforcement Learning*. Phd, Department of Computer Science, University of Massachusetts Amherst.

Sutton, R. (1988). Learning to predict by methods of temporal differences. *Mach. Learn.* 3, 9–44. doi: 10.1007/BF00115009

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

Thirkettle, M., Walton, T., Redgrave, P., Gurney, K., and Stafford, T. (2013a). No learning where to go without first knowing where you're coming from : action discovery is trajectory, not endpoint based. *Front. Psychol.* 4:638. doi: 10.3389/fpsyg.2013.00638

Thirkettle, M., Walton, T., Shah, A., Gurney, K., Redgrave, P., and Stafford, T. (2013b). The path to learning: action acquisition is imparied when visual reinforcement signals must first access cortex. *Behav. Brain Res.* 243, 267–272. doi: 10.1016/j.bbr.2013.01.023

Thorndike, E. (1911). *Animal Intelligence: Experimental Studies*. New York, NY: Macmillan. doi: 10.5962/bhl.title.55072

Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behav. Brain Res.* 199, 119–128. doi: 10.1016/j.bbr.2008.10.030

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.

Received: 17 May 2013; accepted: 23 January 2014; published online: 11 February 2014.

Citation: Shah A and Gurney KN (2014) Emergent structured transition from variation to repetition in a biologically-plausible model of learning in basal ganglia. Front. Psychol. ***5****:91. doi: 10.3389/fpsyg.2014.00091*

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