Biologically Inspired Cognitive Architectures (2014) 10, 10–16

Available at www.sciencedirect.com

ScienceDirect

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

Application of a neural network model of prefrontal cortex to emulate human probability matching behavior

Suhas E. Chelian [a,*], Ryan M. Uhlenbrock [a], Seth Herd [b], Rajan Bhattacharyya [a]

[a] HRL Laboratories, LLC, Malibu, USA
[b] e-Cortex, Boulder, USA

Received 7 November 2014; accepted 7 November 2014

KEYWORDS
Probability matching;
Prefrontal cortex;
n-arm bandit

Abstract
Probability matching behavior occurs in a variety of decision-making domains that can be mapped to the *n*-arm bandit problem. Prefrontal cortex has been implicated in executive control over several tasks including the *n*-arm bandit problem. Previously the Prefrontal cortex Basal Ganglia Working Memory (PBWM) model has been used to replicate other decision-making functions of prefrontal cortex such as recognizing sequences of symbols or visual scenes. In this work, we emulate probability matching behavior from human subjects using the PBWM model in *n*-arm bandit-like problems. Possible extensions to the current work such as including other biases like loss aversion and misperception of both large gains and losses are also discussed.

Introduction

The *n*-arm bandit problem—a class of problems where one must repeatedly chose among several alternatives with unknown and possibly dynamic payoffs—arises in several psychological and technological domains (see Lee, Zhang, Munro, & Steyvers, 2011 for a review). The Bayesian optimal solution is to always pick the option with the highest expected payoff. However, humans often choose options in proportion to the expected payoff of each alternative; this is known as probability matching. A number of neuroscience studies have implicated several regions of prefrontal cortex in the *n*-arm bandit task in rats (Sul, Kim, Huh, Lee, & Jung, 2010), monkeys (Walton, Behrens, Buckley, Rudebeck, & Rushworth, 2010) and humans (Wunderlich, Rangel, & O'Doherty, 2009). A recent model of prefrontal

[*] Corresponding author.
E-mail addresses: sechelian@hrl.com (S.E. Chelian), rmuhlenbrock@hrl.com (R.M. Uhlenbrock), seth.herd@e-cortex.com (S. Herd), rbhattac@hrl.com (R. Bhattacharyya).

http://dx.doi.org/10.1016/j.bica.2014.11.002
2212-683X/© 2014 Elsevier B.V. All rights reserved.

cortex, Prefrontal cortex Basal Ganglia Working Memory, or PBWM, has been used to recognize sequences of symbols (O'Reilly and Frank, 2006) and perform visual scene recognition (Chelian, Bhattacharyya, & O'Reilly, 2011). Here, we adapted PBWM networks to *n*-arm bandit tasks derived from a geospatial intelligence setting where one must choose to defend (e.g., arm 1) or not defend (e.g., arm 2) against an opponent. Greater degrees of conservatism with greater probability matching bias make agents pick options closer to an even distribution than the rational winner-take-all distribution of decisions. E.g., a conservative agent might select the action with the highest expected payoff (e.g., not defend) 60% of the time versus 40% of the time for the other action (e.g., defend). Conversely, lesser degrees of conservatism and probability matching bias correspond to less distance from the optimal distribution of choices; i.e., select the action with the highest expected payoff (e.g., not defend) 90% of the time. These varying degrees of conservatism were found in human data and modeled with PBWM networks.

Materials and methods

We summarize the tasks here but full details can be found in the MITRE Technical Report (in preparation). The tasks were adversarial games set in a geospatial intelligence context with two players, blue and red. Blue was controlled by a human or a neurocognitive model agent while red was a computer opponent. Blue agents received information through various sources of intelligence, or INTs, about red's potential actions. In each trial, blue was informed of the strategic utility (*U*) of a potential attack location and the probability (*P*) of winning a showdown there. Red chose to attack based on *P* and *U* and blue estimated the probability of red attack (*Pa*) with the INTs. Given, *Pa*, *P*, and *U*, blue decided to divert or not divert against a potential attack; we refer to this decision as *D*/∼*D*. In the event that blue did not divert (or did divert) a potential attack and red did attack (or did not attack), no points are lost for either side. If blue diverted and red did not attack, blue has unnecessarily committed resources and loses a small amount of points. If blue did not divert and red attacked, the winner was decided probabilistically using *P* and the winner was awarded *U* points. This is summarized in the payoff Table 1: *U* was 2, 3, 4, or 5. *P* was a real value between 0 and 0.5, and *Pa* was a real value between 0 and 1.

Table 1 Payoff table of the *n*-arm bandit-like tasks.

	Red attacked	Red did not attack
Blue diverted	0	−1
Blue did not divert	±U	0

The optimal strategy is to take the action with the highest expected payoff. From a rational basis, the decision to divert can be decided using the inequality: $-1 + Pa > U \cdot Pa \cdot (2 \cdot P - 1)$. This inequality defines a decision boundary in *Pa*, *P*, *U* space which is illustrated in Fig. 1 for *U* = 2 and *U* = 5. As *Pa* increases along the horizontal axis (or *P* decreases along the vertical axis), red is more likely to attack (or less likely to win a showdown) and hence blue should divert. As *U* increases from Fig. 1a to b at the same *Pa*, *P* point, potential losses increase so blue should also divert. In short, for points below (or on or above) the curve, blue should divert (or not divert) a potential attack.

Blue agents played 5 variations or missions of the geospatial intelligence task. In the first mission, blue practiced estimating *Pa* given *P* and *U* and did not have to make a divert/not divert (*D*/∼*D*) decision. In missions 2 through 5, blue agents made the *D*/∼*D* decision with: mission 2, a basic red opponent; mission 3, a red opponent who could attack in one of two locations but not both; and missions 4 and 5, a red opponent who could vary his strategy in *P*, *U* space in two different ways. All missions had 10 trials except 4 and 5 which had 30 and 40 trials respectively. Data from humans was collected in two rounds, first with 20 subjects (subjects 1—20), then with 30 different subjects (subjects 21—50).

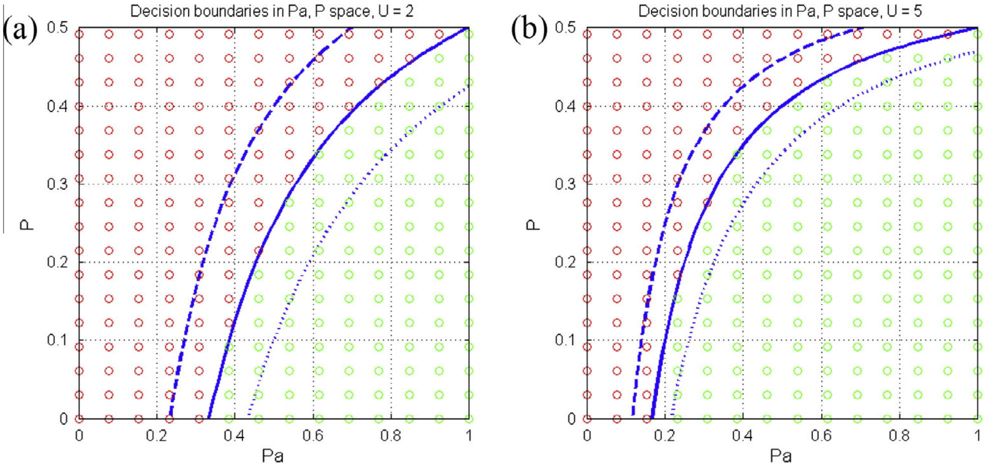

Fig. 1 Decision boundaries in *Pa*, *P*, *U* space for (a) *U* = 2 and (b) *U* = 5. For points below (or on or above) the curve, blue agents should divert (or not divert) a potential attack to minimize expected losses. Diverts and not diverts are depicted as a green or red circles in that order. The rational decision boundary is depicted with a solid line. An aggressive (or conservative) decision boundary with fewer (or more) diverts is depicted with a dotted (or dashed) line; see Section 'Human subjects' for more details.

Each set of subjects had different mission inputs. We built our models using the first set of human subjects and then tested it on the second set of human subjects. Thus to perform well, our models should generalize over human subjects and mission inputs.

Human subjects

Looking at the data from the first 20 human subjects, we found a diversity of responses. We grouped subjects based on the total number of diverts they did on mission 2. This was meant to capture different degrees of conservatism before any significant learning in the task could take place. First we took the average number of diverts across all subjects. Then we grouped those subjects who diverted within one trial on either side of the average; this group was called the moderate group or phenotype. Then we defined the two extremes of the distribution: those who diverted less than the moderate group were called the aggressive group and conversely those who diverted more than the moderate group were labeled the conservative group. Thus the aggressive (or conservative) group diverted less (or more) often than the moderate group. There were 4 (20%), 12 (60%), and 4 (20%) members in the aggressive, moderate, and conservative groups respectively as illustrated in Fig. 2.

We hypothesized that moderate subjects, those closest to the average human behavior, reflect a ''wisdom of the crowds'' and thus they used a decision boundary which was close to rational. We then hypothesized that aggressive subjects shifted their decision boundaries down, performing fewer diverts for the same *Pa*, *P*, *U*, point; conversely, conservative subjects shifted their decision boundary up, performing more diverts for the same point. The aggressive and conservative decisions can be computed by adding a bias term to the rational inequality: $-1 + Pa + \text{bias} > U \cdot Pa \cdot (2 \cdot P - 1)$. A negative bias creates an aggressive phenotype, while a positive bias creates a conservative phenotype; the moderate phenotype uses no bias. E.g., referring back to Fig. 1a, with $Pa = 0.5$, $P = 0.25$, and $U = 2$, a moderate subject would not divert because that point is on the decision boundary ($-1 + 0.5 \ >? \ 2 \cdot 0.5 \cdot (2 \cdot 0.25 - 1)$ or $-0.5 = -0.5$). With an aggressive bias of -0.3, no divert would be made ($-0.5 + -0.3 >? -0.5$ or $-0.8 < -0.5$) but with a conservative bias of +0.3, a divert would be made ($-0.5 + 0.3 >? -0.5$ or $-0.2 > -0.5$). To visualize *D*/∼*D* decisions, then, one can plot whether a human diverted or not for points in *Pa*, *P*, *U* space. A moderate subject should be approximately rational with diverts (or not diverts) below (or on or above) the rational decision boundary. Aggressive and conservative subjects should show similar patterns of diverts and not diverts when compared to their respective decision boundaries which are illustrated in Fig. 1 as dotted and dashed lines in that order. Confusion matrices between divert/not divert behavior and ground truth responses provide complementary characterizations of human subject behavior. For example, the conservative group should have more false positives—diverting when not necessary—than the other groups.

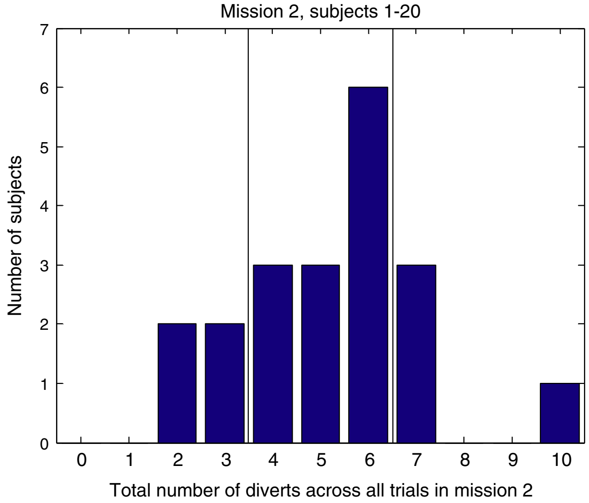

Fig. 2 Histogram of the total number of diverts across all trials in mission 2 for subjects 1—20. Subjects who diverted less than the moderate group (between 0 and 3 times) were labeled as aggressive, while those who diverted more (between 7 and 10 times) were labeled as conservative. The remaining subjects were in the moderate group or phenotype; they diverted between 4 and 6 times.

PBWM networks

To parallel the three groups of humans—aggressive, moderate, and conservative—we trained three different sets of PBWM weights. Training consisted of randomly generating 500 points in *Pa*, *P*, *U* space with *D*/∼*D* decided by the decision boundaries described in Section 'Human subjects' (Divert was encoded as ⟨1, 0, 0⟩ and not divert as ⟨0, 1, 0⟩; this output encoding is arbitrary). Fig. 3a shows a PBWM network with *Pa*, *P*, *U* inputs in the upper left and the *D*/∼*D* decision in the upper right. For 50 epochs, 50 points were randomly chosen from the original set of points and presented to each network. Training converged quickly as shown in Fig. 3b. PBWM networks were then integrated into a larger neurocognitive model to perform other aspects of the task such as receiving INT layers for *P* and *U*. *Pa*, an input to the divert network, is computed by other parts of the network based on intelligence sources and experience on previous trials. The *D*/∼*D* decision and opponent decisions determined the payoff (see Table 1), which provided feedback to other parts of the network. There were a total of 40 instances of the larger neurocognitive model that included the 3 different set of PBWM weights in the approximately the same proportion as the three groups of humans in subjects 1—20.

Comparison between human subjects and PBWM networks

If PBWM networks can emulate performance across human subjects and mission inputs, qualitative trends across the different phenotypes should occur. For example, the number of divert decisions as well as false positives should increase from the aggressive to the conservative phenotypes for human subjects and PBWM networks alike. These trends can be examined by producing *D*/∼*D* decision plots and confusion matrices. In addition, on a trial by trial basis, the average human and PBWM network response can be

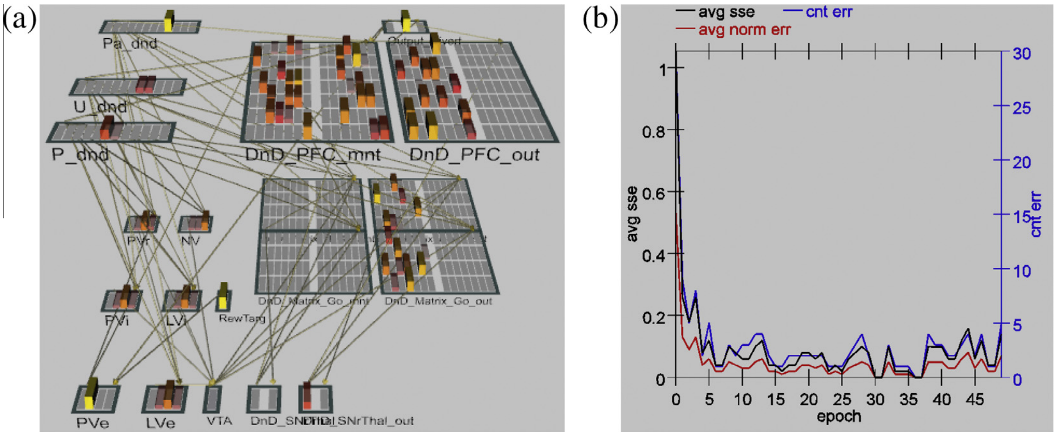

Fig. 3 (a) PBWM networks mapped inputs in *Pa*, *P*, *U* space onto divert/not divert decisions using the three decision boundaries—aggressive, moderate, and conservative—described in Section 'Human subjects'. (b) Training curves for PBWM networks for the conservative decision boundary; training curves for the other decision boundaries were similar.

compared. On trials where humans diverted more than 50% of the time, PBWM networks should do the same (and similarly for trials with less than 50% of divert choices). This similarity can be quantified by correlation coefficients.

Results

Using the definition of groups derived from subjects 1—20, we assigned each subject from subjects 21—50 into the aggressive, moderate, and conservative groups. For example, if a subject diverted 4 (or 7) times in mission 2, they would fall into the moderate (or conservative) group definition (see Fig. 2). For subjects 21—50, we found 2 (6.67%), 16 (53.33%), and 12 (40%) subjects in the aggressive, moderate, and conservative groups in that order. With respect to the first set of subjects, subjects 21—50 have a rightward skew towards more diverts. Below, for each group in subjects 21—50, we plot $D/{\sim}D$ decisions across all missions and decision boundaries using a representative U value of 2.

On the far left of Fig. 4, the aggressive group has more not diverts (red points) than diverts (green points) and the not divert points fall above the aggressive decision boundary with divert decisions mostly falling below it. Furthermore, with respect to the rational decision boundary, the two false positives (green points above the optimal curve) are seemingly random but there is one false negative point (a red point below the optimal curve) which is indicative of an aggressive stance. In the middle of Fig. 4, for the moderate group, there appear to be roughly the same number of not diverts as diverts with more not diverts above the optimal curve than below it and conversely for diverts. The number of false negatives appears to be smaller than the number of false positive but with no apparent pattern. On the far right of Fig. 4, the conservative group has more not diverts than diverts. However, not diverts typically fall above the conservative decision boundary and most diverts fall below it. The confusion matrices shown in Table 2 are also consistent with this trend across groups for subjects 21—50. The number of diverts—the sum of the second column—increases from left to right, from the aggressive to the conservative groups. In addition, so does the number of false positives which is the number in the upper right of each confusion matrix. Interestingly there is no general trend with percent correct—85%, 76.5% and 81.46% for aggressive, moderate and conservative groups in that order for subjects 21—50.

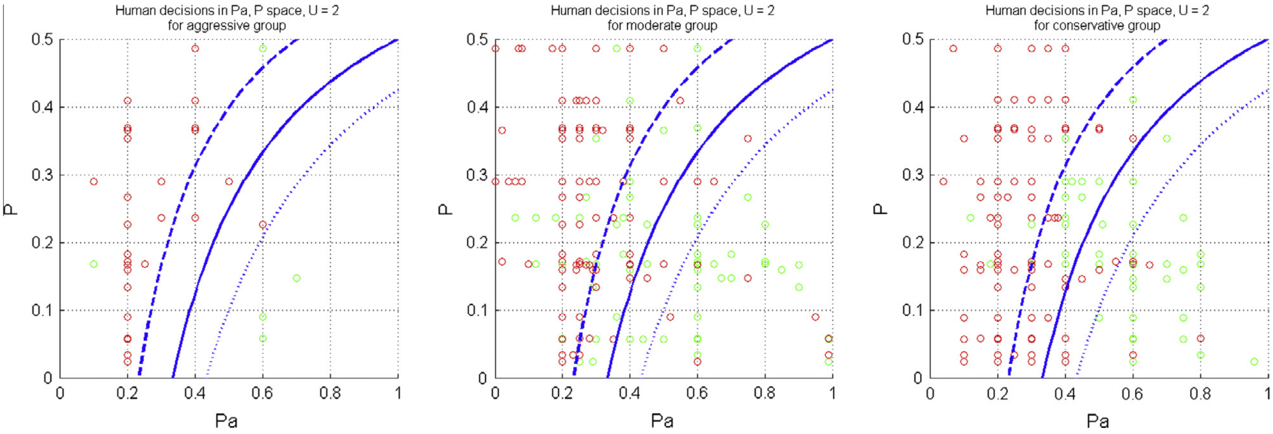

Fig. 4 Divert/not divert decisions across all missions for human subjects 21—50, from left to right, aggressive, moderate and conservative groups. The same bounds and decision boundaries shown in Fig. 1a are also shown here (U = 2).

Table 2 Confusion matrices for human subjects 21—50 for (a) aggressive, (b) moderate, and (c) conservative groups; each confusion matrix is an average over all subjects in a group. In each confusion matrix, rows are the ground truth and columns are the responses. (For example, a false positive response of diverting when there is no attack would be in the 1st row, 2nd column.)

	Subjects 21—50							
	~D	D		~D	D		~D	D
	(a) Aggressive			(b) Moderate			(c) Conservative	
~D	63.75	12.50	~D	50.62	14.53	~D	52.08	16.04
D	2.50	21.25	D	8.91	25.94	D	2.50	29.38

PBWM networks

Next we examined the decisions made by our PBWM networks to see if they generalized to new data points and matched different phenotypes of human decision making.

For all groups, other parts of the larger neurocognitive network estimated *Pa* lower than humans. This shifted points in *Pa*, *P* space to the left, above even the aggressive decision boundary causing more diverts than humans did. On the far left of Fig. 5, the aggressive set of models has fewer diverts than not diverts. Furthermore, with respect to the optimal decision boundary, there are more false negatives (3) than false positives (0), indicative of an aggressive stance. In the middle of Fig. 5, for the moderate set of networks, diverts fall below the optimal decision boundary while not diverts fall above it. There are neither false positives nor false negatives. On the far right of Fig. 5, the conservative set of networks produces more diverts than not diverts with diverts below the conservative decision boundary and conversely for not diverts (with the exception of 2 false negatives). Furthermore, with respect to the optimal decision boundary, there are more false positives than false negatives indicative of a conservative stance. The confusion matrices shown in Table 3 are also consistent with this trend across groups for model runs for subjects 21—50. The number of diverts—the sum of the second column—increases from left to right, from the aggressive to the conservative groups. In addition, so does the number of false positives which is the number in the upper right of each confusion matrix. These trends match the human data. Here for the model runs for subjects 21—50, there is general trend with percent correct—90.5%, 97% and 93%

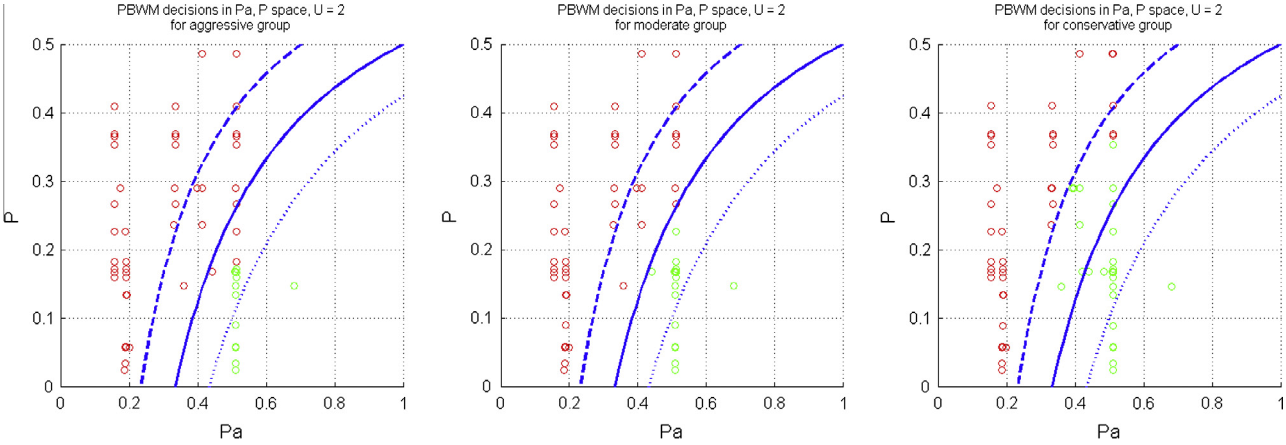

Fig. 5 Divert/not divert decisions across all missions for PBWM model runs for human subjects 21—50, from left to right, aggressive, moderate and conservative groups. The same bounds and decision boundaries shown in Fig. 1a are also shown here ($U = 2$).

Table 3 Confusion matrices for PBWM networks for subjects 21—50 for (a) aggressive, (b) moderate, and (c) conservative groups; each confusion matrix is an average over all model runs in a group. In each confusion matrix, rows are the ground truth and columns are the responses. (For example, a false positive response of diverting when there is no attack would be in the 1st row, 2nd column.)

	PBWM for subjects 21—50							
	~D	*D*		*~D*	*D*		*~D*	*D*
	(a) Aggressive			*(b) Moderate*			*(c) Conservative*	
~D	65.50	0.00	*~D*	63.00	3.00	*~D*	59.00	7.00
D	9.50	25.00	*D*	0.00	34.00	*D*	0.00	34.00

for aggressive, moderate and conservative groups in that order meaning the moderate group was most rational.

Comparison between human subjects and PBWM networks

To compare trial by trial responses, we now show a time series comparing the average human response to the average PBWM response on mission 2 for subjects 21—50 in Fig. 6 (again, mission 2 is the mission with the least complexity). When more than 50% of humans divert, PBWM networks do likewise in trials 2, 3, 4, 7, 9, and 10. Similarly when less than or equal to 50% of humans divert in trials 1, 5, 6, and 8, PBWM networks also divert less than 50% of the time with the exception of trial 8 (trials 1 and 6 also show much less diverts from PBWM).

To quantify the agreement between average human and average PBWM network responses, we list the trial by trial correlation coefficients between humans and PBWM networks for each mission for subjects 21—50 in Table 4. Across all missions, correlations are high and statistically different from chance except for mission 3. On mission 3, the complication of 2 possible attack points may have led to a lower correlation coefficient.

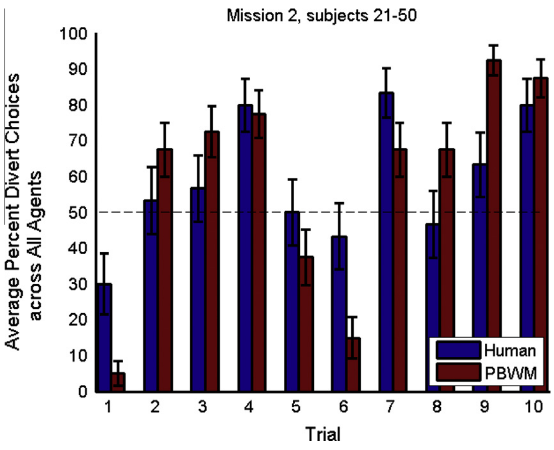

Fig. 6 Comparison of average percentage divert choices on each trial between humans (red) and PBWM networks (blue) on mission 2 for the subjects 21—50. Generally when humans divert more than 50% of the time so does PBWM and similarly for less than or equal to 50% of the time.

Table 4 Trial by trial correlation coefficients between humans and PBWM networks for each mission. A star represents a statistically significant correlation (*t*-test, *p* < .05).

Mission	Subjects 21—50
2	0.76*
3	0.57
4	0.77*
5	0.76*
Average	0.72

Discussion and conclusion

Here we have shown that the PBWM models of decision making can model different phenotypes of human behavior even with novel trial stimuli and subjects. For both humans and PBWM networks, conservative agents showed more probability matching bias than aggressive or moderate agents, with more false positives and a larger number of divert decisions than the other two groups.

Herd, Krueger, Kriete, Huang and O'Reilly (2013) describe an extension to PBWM, called PBDM (Prefrontal cortex Basal ganglia Decision Making). PBDM extends previous theories of how frontal cortex and basal ganglia work together to perform action selection (Frank, Seeberger, & O'Reilly, 2004) and selection of items to maintain in working memory (O'Reilly and Frank, 2006) to account for more abstract types of human decision making. In this particular task with two possible choices, divert or not divert, the selection of candidate actions is trivial, and we assume that the two options are weighed in parallel, in separate sets of ''stripes'' which are continuous, related circuits through PFC and BG. PBDM can either weigh a few well-learned options in parallel, or consider a single more novel option or context at a time, serially. Our choice of treating this task as well-learned is somewhat at odds with the relative novelty of the task, but good fits to human data suggest that it is not unreasonable—the decision framing of ''make a bet or not'' could be considered well-learned if it is successfully generalized from the many related experiences humans encounter in their lives (e.g., cross the street or not when faced with the Walk/Don't Walk sign, bet or not bet in casino games, etc.).

The model's qualitative and quantitative fits to human data are encouraging. It also fits a broader range of functional and physiological data than more abstract functional models of human decision making. This work is some of the first to apply the theories of PFC and BG function to abstract decision making, and so much more work is needed to evaluate its ability to more broadly match human performance. One such topic is the phenomena of risk aversion and loss aversion, biases seen in different individuals and different contexts. Applying our model of reward-predictive dopamine release (Hazy, Frank, & O'Reilly, 2010) shows promise in explaining these complex effects. Current work is also seeking to address the contribution of these same mechanisms to motivated reasoning effects, in which social and other types of rewards influence reward-directed decision making in the realm of evaluating complex evidence to arrive at beliefs.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) Contract Number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be

interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

Chelian, S.E., Bhattacharyya, R., & O'Reilly, R. (2011). Learning categories with invariances in a neural network model of prefrontal cortex. In *Proceedings of the second annual meeting of the BICA society (BICA 2011), Arlington, USA*.

Frank, M. J., Seeberger, L., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science, 306*, 1940—1943.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. *Neuroscience and Biobehavioral Reviews, 34*, 701—720.

Herd, S. A., Krueger, K. A., Kriete, T. E., Huang, T., & O'Reilly, R. C. (2013). Strategic cognitive sequencing: A computational cognitive neuroscience approach. *Computational Intelligence and Neuroscience*, 149329.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research, 12*, 164—174.

MITRE Technical Report (in preparation). *Integrated cognitive-neuroscience architectures for understanding sensemaking (ICArUS): phase 2 challenge problem design and test specification*.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the frontal cortex and basal ganglia. *Neural Computation, 18*, 283—328.

Sul, J. H., Kim, H., Huh, N., Lee, D., & Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron, 66*, 449—460.

Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron, 65*, 927—939.

Wunderlich, K., Rangel, A., & O'Doherty, J. P. (2009). Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences, 106*, 17199—17204.