On Modeling Human Learning in Sequential Games with Delayed Reinforcements

Roi Ceren, Prashant Doshi
Department of Computer Science
University of Georgia
Athens, GA 30602
{ceren,pdoshi}@cs.uga.edu

Matthew Meisel, Adam Goodie
Department of Psychology
University of Georgia
Athens, GA 30602
{mameisel,goodie}@uga.edu

Dan Hall
Department of Statistics
University of Georgia
Athens, GA 30602
danhall@uga.edu

***Abstract*—We model human learning in a repeated and sequential game context that provides delayed reinforcements. Our context is significantly more complex than previous work in behavioral game theory, which has predominantly focused on repeated single-shot games where the actions of other agent are perfectly observable and provides for an immediate reinforcement. In this complex context, we explore several established reinforcement learning models including temporal difference learning, SARSA and Q-learning. We generalize the default models by introducing behavioral factors that are reflective of the cognitive biases observed in human play. We evaluate the model on data gathered from new experiments involving human participants making judgments under uncertainty in a repeated strategic and sequential game. We analyze the descriptive models against their default counterparts and show that modeling human aspects in reinforcement learning significantly improves predictive capabilities. This is useful in open and mixed networks of agent and human decision makers.**

***Keywords*—cognitive science, multi-agent systems, probability judgment, reinforcement learning**

I. Introduction

We study the computational modeling of human learning in a repeated, strategic sequential game context that exhibits delayed reinforcements. This context is substantially more complex than in previous related work [1], [2], [3], [4], which predominantly models learning strategies in repeated single-shot games. In these latter games, actions of the other agent are perfectly observable and provide for an *immediate* observation and payoff reinforcement.

We conducted experiments involving human subjects playing a strategic, sequential game repeatedly. Participants observed an unmanned aerial vehicle (UAV) move through a theater and assessed the chance of reaching a goal sector without being spotted by a hostile UAV at a series of steps. The other UAV's movement is fixed but not revealed. Data collected from these new experiments help us in understanding, in part, whether the judgments over a series of games reflect *learning*. In this complex context, we observe remarkable learning and provide a behavioral model of aggregate learning.

Because, (a) the context is sequential; (b) positive or negative delayed reinforcements in the form of safely reaching the goal sector thereby collecting a large reward or being spotted and not obtaining a reward, respectively, are obtained; and (c) the other agent remains hidden from the participant except when the participant's UAV is spotted, we adopt a sequential reinforcement learning model. This model is more general compared to the previous models involving cumulative attractions [2], [3].

In the experiments, humans provide likelihood assessments only and do not control the UAV in order to avoid a cognitive overload. The UAV moves along various predefined trajectories. These serve as exploration policies, and motivates using *on-policy* reinforcement learning methods such as TD(λ) [5] and SARSA [6]. We include Q-learning [7] as well for completeness.

TD models the desirability of a state based on previous experiences in that state and next states in the trajectory. TD may be extended by propagating rewards backwards to previously visited states using eligibility traces. Bogacz et al. [8] demonstrate the utility of considering short eligibility traces in TD toward modeling human actions in a sequential economic game, which motivates its inclusion here. SARSA [6] expands the model by including the actions performed in the current and next state. Finally, Q-learning alters the model by choosing the best possible action at the next state.

This paper contributes *process models* that use cognitive insights for modeling new data on human learning in repeated, sequential games. These models differ from statistical curve fitting such as regression analysis and kernel-based density estimation on the on the data by providing some insights into the judgment and decision-making processes that potentially led to the observed data. Human behavior often does not adhere to normative prescriptions [9]. Rewards or penalties in a state are perceived to *spill over* to neighboring states, thereby affecting the play in those states [10]. These cognitive biases may influence reinforcement learning. We generalize the default models using behavioral parameters.

Because participants assess probabilities, their assessments may be susceptible to *subproportional* probability weighting [11], [12], [13] – a phenomenon involving under- or over-statement of probabilities. Consequently, we integrate weighting functions that subproportionally map the normalized values to reported assessments. Evaluation of the extended models demonstrates a significantly improved fit over the default models. Simulations of the models demonstrate an effective fit of the data but also reveal room for improvements.

978-1-4799-0652-9/13/$31.00 ©2013 IEEE

In addition to insights into the cognitive process by which humans act in strategic settings, the predictive models are useful in open and mixed agent-human networks. For example, assistive agents in learning environments may utilize the behavioral parameters to better model student actions.

II. Experiments on Assessing Probabilities

In an IRB-approved study conducted with human participants, we collected probability assessments elicited at various points in a strategic, uncertain two-agent game.

A. Setting: UAV Game

In order to evaluate probability judgments of human operators of UAVs, we formulated a strategic game involving uncertainty. In this sequential game, participants observe a UAV moving through a 4×4 theater of sectors. From an initial sector, the UAV moves towards a goal sector (represented by the shaded sector $\langle 3, 2 \rangle$), as we show in Fig. 1, using different predefined trajectories. The environment is shared with another hostile UAV starting in sector $\langle 2, 3 \rangle$. While participants are briefed about the starting sector of the other UAV and that it moves in a loop, no other information, such as the movement or actions of the other UAV, is revealed to them while playing the game. Unknown to the participants, the other UAV's trajectory is fixed and both UAVs move synchronously.

A trial representing the completion of a trajectory by the participant's UAV is considered a win if the participant's UAV reaches the safe goal sector, or a loss if it is *spotted* by the other UAV. The participant's UAV is spotted when both UAVs come to occupy the same sector, after which the trial ends.

B. Methodology

Participants play a total of 20 trials. Within each trial, a participant encounters a series of assessment points along a predefined trajectory of the participant's UAV. The UAV trajectories vary with the trials. At each decision point, participants are asked to fill a questionnaire and assess the probability of reaching the goal sector safely without being spotted. In Fig. 1, we illustrate the first two decision points within an example trial. Notice that the participant is shown the entire trajectory that her UAV will travel to facilitate an informed judgment.

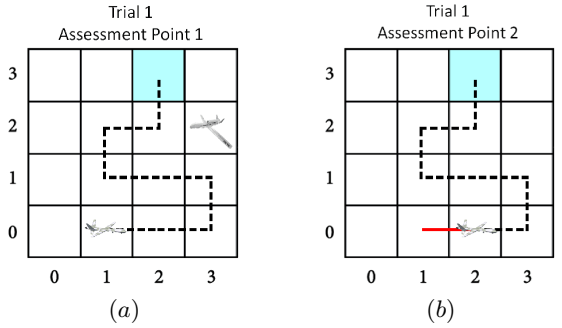

Fig. 1. First two assessment points of a particular trial. (a) The other UAV and the trajectory of the participant's UAV are shown at the first decision point. This confirms the presence of another UAV in the theater for the participant. (b) Traveled trajectory is highlighted and the other UAV is no longer shown.

C. Results

A total of 43 participants experienced the game. In order to analyze for evidence of learning, we performed a trend analysis utilizing a generalized linear regression model. The model was designed to determine the mean slope of the assessments in a trial averaged across all trials and participants, and determine the gradient of the changes in the mean slope as the trials progressed averaged over all participants. The assessments are modeled as changing linearly across the assessment points and the mean slope is modeled as changing linearly as the trials progress, with a random intercept for initial probability assessments. We model the participant anticipating a win or a loss as having an effect on their assessments, which is justified by the significant difference in values of the variables when analyzing data on wins and losses separately.

Statistic	**Losses**		**Wins**	
	Estimate	**S.E.**	**Estimate**	**S.E.**
intercept	0.3315	0.032	0.5392	0.03
slope within trial	0.02053	0.006	0.05395	0.005
slope grad. bet. trials	-0.00486	0.001	-0.00129	0.000

TABLE I. Mixed effect linear regression on probability judgments of participants separated by wins and losses in the trials. All values are significant, $p \ll 0.01$.

In Table I, we show the results of the statistical analysis. Notice the positive intercepts for both wins and losses and the very small standard errors (S.E.) with $p \ll 0.01$ indicating a significant fit. More importantly, the mean slope within a trial is positive and a significant p-value indicates that participant probabilities increase as they approach the goal. Furthermore, the negative value for the mean slope gradient between trials indicates that the mean slope reduces as the trials progress, and that this reduction is significant.

The positive mean slope indicates that participants generally demonstrate greater certainty as reflected in their increasing probabilities of reaching the goal without being spotted, as a trial progresses and they get closer to the goal. This remains true for losses as well although the mean slope is substantially smaller compared to that for wins.

Importantly, the negative gradient in mean slope between trials indicates that participants are not changing their probability assessments in a trial as much as they were in previous trials. For the ideal case where participants precisely know how the other UAV is moving, they would be certain about the outcome given their trajectory and their assessments would not vary within a trial. Therefore, a reducing change in the judgments is indicative of participants gradually demonstrating greater confidence in their assessments. We interpret these results as indicative of learning from previous experiences.

III. Descriptive Reinforcement Learning

Learning may be defined as an observed change in behavior due to experience [9]. The relative growth in confidence as the trials progress and generally increasing probability assessments within a sequential game provide evidence of participants learning as they repeatedly play the sequential game. Being spotted in a sector provides a negative reinforcement for being in that sector at that time while reaching the target results in

a strong positive reinforcement. Consequently, we hypothesize that reinforcement learning could provide an explanation for the data in general.

A. Default Models

TD(λ), SARSA and Q-learning represent three different ways of implementing the temporal difference error in reinforcement learning. The use of predefined UAV trajectories makes on-policy TD(λ) and SARSA appropriate although Q-learning may not be definitively ruled out.

1) Temporal Difference (TD(λ)): TD(λ) computes a value that signifies the desirability of being in a state [5]. At each step, TD alters the previous value of the state by adding the immediate and potential future reward decayed by a discount parameter, $\gamma \in (0, 1]$. The effect of the reward is mediated by a learning parameter, $\alpha \in [0, 1]$. Equation 1 formalizes TD(λ):

$$V(s; \alpha, \lambda) = V(s) + \alpha\big(r(s) + \gamma \cdot V(s') - V(s)\big)e(s; \lambda) \quad (1)$$

where s and s' denote the current and next state due to a given action, respectively. We model a state to consist of the UAV's sector and the time step (the number of moves so far) initially 0, s = ⟨sector, timestep⟩. Time step is included to model the fact that the desirability of a sector may vary dynamically based on the time elapsed because of the presence of the other UAV moving in the theater. The time step is bounded by the length of the longest trajectory. Immediate reward, $r(s)$, is -1 for the state where the participant's UAV is spotted and 1 for the state when the goal sector is reached. For any other state, no immediate reward (reinforcement) is obtained. Values of all states-action pairs other than the states containing the goal sector are initialized to 0. For the states with the goal sector, the value is initialized to 1.

Eligibility traces represent a short-term memory of the visited states and provide a way to implement a discounted look ahead of more than one step. Their inclusion has been known to speed up the learning. Previously visited states are eligible to receive a portion of the temporal difference error as guided by Eq. 2.

$$e(s; \lambda) = \begin{cases} \gamma\lambda e_{t-1}(\hat{s}), & \text{if } \hat{s} \neq s \\ \gamma\lambda e_{t-1}(\hat{s}) + 1, & \text{if } \hat{s} = s \end{cases} \quad (2)$$

Here, $e(s; \lambda)$ is initialized to 0 for each state. Eligibility traces are precluded if $\lambda = 0$. In our application, we may initialize $e(s; \lambda)$ to 0 after each trial so that the credit is not assigned across trajectories.

2) SARSA: SARSA extends on-policy TD by assigning a value to the combination of state and action. While the state is as defined previously for TD, the actions include the participant's UAV moving to the north, east, and west. Let the action given by a trajectory from a state, s, in trial, k, be denoted as $\pi^k(s)$. Equation 3 formalizes the update rule:

$$\begin{aligned} Q(s, \pi^k(s); \alpha) = \quad & Q(s, \pi^k(s)) + \alpha\big(r(s) + \gamma \cdot Q(s', \pi^k(s')) \\ & - Q(s, \pi^k(s))\big) \end{aligned} \quad (3)$$

Q-values of all states-action pairs other than the states containing the goal sector are initialized to 0. For the states with the goal sector and any action, the Q-value is initialized to 1.

SARSA may be extended to include eligibility traces analogously to TD.

3) Q-learning: While TD and SARSA compute the value of a given exploration policy (this is often called the prediction problem in reinforcement learning), Q-learning performs off-policy learning. Specifically, the future reward in the temporal difference error is the maximal Q-value of the next state, s', by choosing the maximizing action that may not be the one on the trajectory.

$$\begin{aligned} Q(s, \pi^k(s); \alpha) = \quad & Q(s, \pi^k(s)) + \alpha\big(r(s) + \\ & \gamma \cdot \max_{a'} Q(s', a') - Q(s, \pi^k(s))\big) \end{aligned} \quad (4)$$

Here, state s is as defined previously and $\pi^k(s)$ denotes the action given by the trajectory in trial k of the experiment.

B. Behavioral Generalizations

Previous applications of reinforcement learning in the context of repeated games demonstrate the plausibility of some cognitive biases affecting humans while playing games [10], [3], [9]. We briefly review previous applications followed by introducing these factors into the models. By modeling human factors observed in game-theoretic experiments within learning, we establish a novel and sophisticated framework for descriptive reinforcement learning.

Erev and Roth [3] model the *attraction* toward a strategy c at time step t, $A_c(t)$, as the sum of the previous attraction to strategy c and its immediate reward. An attraction represents the desirability of taking an action, and is analogous to the Q-value with the difference that it is not bounded. A Q-value is bounded by $\frac{R_{max}}{1-\gamma}$, where R_{max} is the maximum possible reward. Reinforcement learning in repeated games takes the form, $A_c(t) = A_c(t-1) + r$. The absence of a state variable is due to the model being usually applied to repeated games.

Insights from applying these models to behavioral data reveal phenomena that better explain the non-normative nature of human learning [3], [14]. Such factors include *forgetfulness* – previous information is degraded in its effect – *spill over* – partially attributing reward or penalty to neighboring strategies or states [1] – and *subproportional weighting* – the phenomenon by which humans under- or over-weight their judgments. These lead to behavioral generalizations of reinforcement learning, which potentially better model the data.

1) Forgetfulness and Spill Over: Forgetfulness implies that experience from a previous trial has a diminished effect on current assessments. Observe that $1-\alpha$ in the previous models where learning rate, $\alpha \in [0, 1]$, discounts the desirability of the state or state-action whose value is being updated as computed from previous trials. Consequently, parameter α effectively models forgetfulness balancing it with new information.

[1] An illustrative example is that of the roulette player who bets on a particular number only to land on a nearby number [10]. The player may have her guess confirmed though she lost the bet.

Spill over involves the misattribution (or "generalization") of rewards to neighboring strategies. In sequential games, neighboring strategies could be interpreted as neighboring state-action pairs. Consequently, the obtained reward may spill over to nearby sectors occupied in current, preceding or subsequent time steps. Figure 2 illustrates multiple approaches to spill over contextual to our domain. A lightly shaded square indicates the experienced sector. Darker shaded sectors each receive a portion of the spilled over reward equal to ϵ.

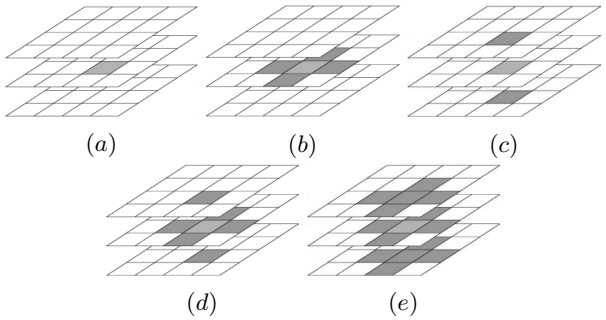

Fig. 2. Differing ways of implementing spill over. (a) Normative case where no other neighbors are attributed a reward from an experienced sector. (b) Adjacent sectors at current time step receive some spill over reward. (c) No neighbors at the current time step receive a spill over but the visited sector in the preceding and following time steps are attributed some reward. (d) Merges spill overs of (b) and (c). (e) Adjacent sectors at current, preceding and following time steps receive spill over rewards.

Spill over proportion, $\epsilon \in [0, 1]$, attributes a fraction of the reward obtained at the state, $r(s)$, to neighboring states or state-action pairs. For convenience, we keep the action in the neighboring state-action pairs fixed and the same as in the state-action under consideration. We generalize SARSA (Eq. 3) with spill over as shown below. Its use in TD(λ) and Q-learning is analogous.

$$Q(s, \pi^k(s); \alpha, \epsilon) = \begin{array}{l} Q(s, \pi^k(s)) + \alpha\big((1-\epsilon)r(s) + \\ \gamma \cdot Q(s', \pi^k(s')) - Q(s, \pi^k(s))\big) \end{array} \quad (5)$$

If $\epsilon > 0$, a neighboring state-action pair with the neighboring state denoted as s_n and (for SARSA and Q-learning) same action, $\pi^k(s)$, accumulate a fraction of the reward obtained. Neighboring states are determined by our selection of a spill over implementation from Fig. 2.

$$Q(s_n, \pi^k(s); \alpha, \epsilon) = Q(s_n, \pi^k(s)) + \alpha\big(\epsilon \cdot r(s) + -Q(s_n, \pi^k(s))\big) \quad (6)$$

Equations 5 and 6 represent our behavioral generalizations to SARSA. However, the values must be mapped to probability assessments for modeling the data. Observe that values approaching -1 represent a path likely to lead to a loss and those approaching 1 indicate a win from that path. Because the value is representative of the *desirability* of the state (and action), it maps to the likelihood of success from that state given the trajectory, naturally. We may then convert the values to assessments by normalizing them between 0 and 1. However, humans reporting their judgments do not always follow a linear mapping as we discuss next.

2) Subproportional Weighting: Humans tend to misrepresent their probability judgments in decision-making processes [11], [12], [13]. Prospect theory [11] notes that the weights given to probability assessments and their payoff values are usually not linear. Humans tend to under- or over-weight their probability assessments in domains involving chance [15]. In the UAV game, participants are asked about the probability for overall success in the current trial as it progresses and are compensated based on the assessment, which is susceptible to being under- or over-stated.

Several subproportionality weighting functions exist that map believed probabilities to weighted assessments using a sigmoidal or inverse sigmoidal function. The two-parameter model [13] defines the, (i) curvature of the function, and (ii) elevation of the median weight. Prelec's one-parameter model [12] fixes elevation and allows the curvature, parameterized by β, to vary:

$$w(p; \beta) = e^{-(-ln(p))^{\beta}} \quad (7)$$

For our purposes, p is the normalized value and $w(p)$ is the weight assigned to it. $w(p)$ therefore serves as the output assessment of the weighted model.

IV. Performance Evaluation

Multiple parameters affect the performance of our descriptive models. We learn their values by optimizing the fit of the model to the collected data. Subsequently, we compare the fits of default TD(λ) with λ set to 0 and 1, SARSA, Q-learning, their behavioral generalizations and the null hypothesis as a broad set of models that fit the learning context of the domain. We then isolate the model demonstrating the best fit of the data and evaluate its predictive performance.

A. Learning Behavioral Parameters

Data collected from the 43 participants were randomly partitioned into 5 equal folds. Utilizing the Nelder-Mead method – a downhill simplex method for minimizing an objective function – a model is trained over 4 folds and then tested over the remaining fold.

Beginning at the first time step of the first trial, values are updated as the participant's UAV follows trajectories in the trials. As mentioned in Section III, on being spotted a reward of -1 is obtained for that state. If it reaches the goal sector, a reward of 1 is obtained for the state, otherwise there is no reward. For all the participants in the training folds, we update the function and simultaneously predict probabilities for the 20 trials that each participant experiences. Parameters are learned by minimizing the sum of squared differences (SSD) between the stated probabilities of a participant, pt, at each assessment point, i, in a trial, k, $p_e(i, k, pt)$, and those predicted by our general model, $p_m(i, k)$. We may interpret this difference as the *fit* of the model with smaller differences signifying better fits. Formally,

$$SSD = \sum_{pt=1}^{N} \sum_{k=1}^{20} \sum_{i} (p_m(i, k) - p_e(i, k, pt))^2 \quad (8)$$

where, N is the number of participants in the training folds and i is the number of UAV actions in a trial, which vary between trials.

B. Results

We begin by learning the parameter values and establishing the best fitting spill over among those shown in Fig. 2. We implement each spill over in each of the default models and perform a 5-fold cross validation summing the SSDs over the test folds. Table II lists the SSDs for the different spill over implementations in each model. Notice that each implementation in SARSA provides the best fit among the different learning models. For SARSA, the implementation which spills the reward across adjacent sectors (**local**) results in the best fit.

	TD(0)	TD(1)	SARSA	Q-learning
No Spill over	360.187	407.469	**355.569**	379.463
Local	351.258	382.345	**341.923**	372.306
Time Step	361.141	400.23	**346.425**	371.776
Local & Time Step	356.304	392.01	**354.161**	364.861
All Neighbors	348.165	383.386	**343.458**	358.075

TABLE II. SSDs FOR THE DIFFERENT SPILL OVER IMPLEMENTATIONS SHOWN IN FIG. 2 SUMMED OVER ALL TEST FOLDS. SARSA PROVIDES THE LOWEST SSD FOR EACH SPILL OVER IMPLEMENTATION AND **Local** FITS THE BEST.

Table III shows the learned values of the three behavioral parameters in each descriptive model utilizing the spill over implementation that results in the lowest SSD for that model. We fix the discount factor, γ, to 0.9. A $1-\alpha$ value of 0.421 for SARSA signifies that participants place a moderately lower emphasis on their previous experiences as compared to the current and future reward, thereby forgetting them. We experimented with a linearly varying α as well resulting in a worse fit. On the other hand, the spill over is negligible for SARSA but substantial for the other models. Curvature of the subproportional weighting as parameterized by β remains above 1 for all models indicating that the function is sigmoidal.

	TD(0)	TD(1)	SARSA	Q-learning
α	0.570	0.750	0.579	0.491
ϵ	0.215	0.463	0.0004	0.809
β	1.905	1.785	2.045	1.420
Total SSD	348.165	382.345	341.923	358.075

TABLE III. LEARNED PARAMETER VALUES OF THE DIFFERENT MODELS.

Table IV shows the comparative performance of the different reinforcement learning models including a random model. We observe that the corresponding behavioral generalization improves each default model with Behavioral SARSA showing the lowest SSD and therefore the best fit. It outperforms the next best model (TD(0)) significantly (Student's paired, two-tailed t-test, p-value < 0.05), as well as its default model.

C. Predictive Performance

While SSDs show comparative performance, it does not clearly reveal how well the models predict the observed data. We visually explore the model predictions in this section. We select the best performing model from the previous subsection, Behavioral SARSA, and plot its performance. Because the experiment utilized trajectories of differing lengths for the participant's UAV between the win and loss trials, and the collected data also exhibits difference between the two (for e.g., see Table I), we present the results separately for the two types of trials for clarity. However, a single model was trained over the participant data.

Model	total SSD
Behavioral SARSA	**341.923**
Behavioral TD(0)	348.165
Default SARSA	355.569
Behavioral Q-learning	358.075
Default TD(0)	360.187
Default Q-learning	379.463
Behavioral TD(1)	382.345
Default TD(1)	407.469
Random	891.18

TABLE IV. BEHAVIORAL SARSA SHOWS THE BEST FIT AND THE DIFFERENCES WITH OTHERS ARE SIGNIFICANT.

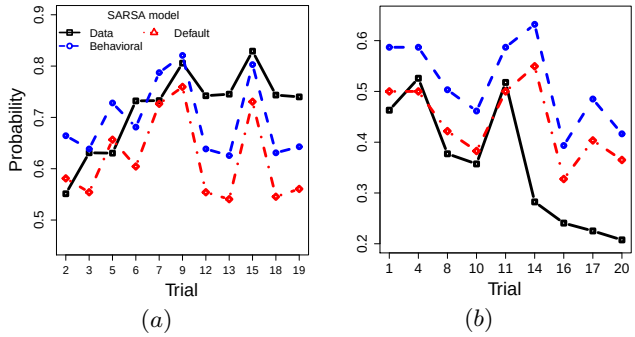

Fig. 3. Average probability assessment in each trial for trajectories that lead to (a) successfully reaching the goal sector, and (b) being spotted by the other UAV.

In Fig. 3, we show the average probability over all assessment points and participants for each trial, separated by winning and losing trials. For wins and losses, model predictions fit the general shape of the probability changes closely. For the win trials, Behavioral SARSA effectively models the changing mean assessments per trial up to and including the last trial. We show the performance of the default SARSA across the trials as well. As Table IV suggests, Behavioral SARSA demonstrates improved predictions across the trials compared to the default.

Trajectories that result in a win are of lengths 4, 6, or 8 time steps. Probability assessments are substantially affected by the distance to the goal sector, so we analyze the data and model predictions separately for each of these lengths in Fig. 4 next. While Behavioral SARSA understates the probabilities in comparison to the data for the longer trajectories, it exhibits the overall trend correctly for the trajectories of different lengths.

In Fig. 5, we compare the model predictions with the data averaged over all the trials that result in a loss. Participants, on average, start with lower assessments compared to trials that result in a win. This indicates that participants are generally good at identifying eventual losses and retain their pessimism as trials progress. The models show higher initial judgments

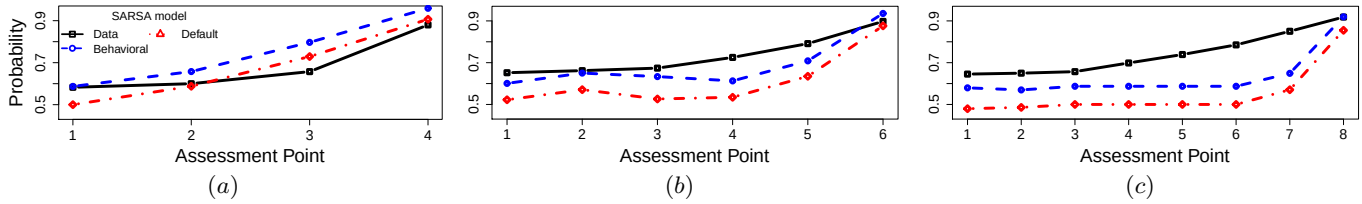

Fig. 4. Comparison of predicted judgments by the different models with the experiment data for trajectory lengths of, (a) 4 time steps, (b) 6 time steps, and (c) 8 time steps. Vertical bars are the standard errors.

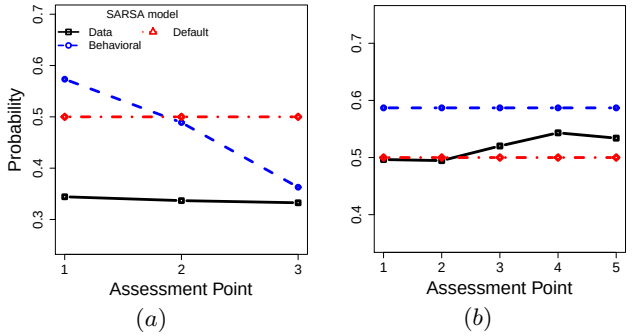

Fig. 5. A comparison of model predictions with observed data for loss trajectories. We show the comparisons for trajectories of lengths (a) 3 time steps, and (b) 5 time steps.

and exaggerated decreases in their probability predictions over time compared to the slight dip in probabilities we observed for trajectories of length 3 (Fig. $5(a)$). For the longer trajectory, participants generally became more optimistic until just before their loss, while the models' predictions averaged over all such trials remain mostly flat. The primary reason is the lack of substantive data because each participant experiences just one trial that results in being spotted after 5 steps.

V. Discussion and Future Work

Descriptive reinforcement learning with cognitive biases gets us close to modeling human judgments in contexts with delayed reinforcements but shows room for further improvement. Certain behaviors are challenging to computationally model, such as participants dropping their assessments in the later stages (Fig. $5(b)$). This observation illuminates a pitfall of reinforcement learning: model assessments may propagate too slowly to precisely match the data set, an observation that has precedence [1]. While participants may quickly change their assessments, temporal difference learning requires several iterations before a dramatic change is visible.

Other learning models may also exemplify these strategic tasks. In this work, it is assumed that the participant does not explicitly model the other UAV's movements but characterizes losses as an environmental event. Belief-based learning models establish beliefs on opponent models in competitive games. An avenue of future work is to model descriptive learning in this domain by including behavioral parameters in a sequential belief-based learning model. The presence of predefined trajectories may also motivate explorations of models such as Markov chains. However, the movements of the participant's UAV are deterministic and very less information about the other is available. Therefore, though Markov chains are useful for learning the distribution over the states, the chain may take a long time to converge.

Acknowledgments

This research was supported by a grant from Army RDECOM, #W911NF-09-1-0464, to Prashant Doshi, Adam Goodie and Dan Hall. We thank Ryan T. Gell for his help with the statistical analyses.

References

[1] P. H. McAllister, "Adaptive approaches to stochastic programming," *Annals of Operations Research*, vol. 30, pp. 45–62, 1991.

[2] A. Roth and I. Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, vol. 8, pp. 164–212, 1995.

[3] ——, "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, vol. 88, no. 4, pp. 848–881, 1998.

[4] I. Erev and G. Barron, "On adaptation, maximization, and reinforcement learning among cognitive strategies." *Psychological Review*, vol. 112, no. 4, pp. 912–931, 2005.

[5] R. S. Sutton, "Learning to predict by the methods of temporal differences," in *Machine Learning*. Kluwer Academic, 1988, pp. 9–44.

[6] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*. University of Cambridge, Dept. of Engineering, 1994.

[7] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[8] R. Bogacz, S. M. McClure, J. Li, J. D. Cohen, and P. R. Montague, "Short-term memory traces for action bias in human reinforcement learning," *Brain Research*, vol. 1153, pp. 111–121, 2007.

[9] C. Camerer, *Behavioral Game Theory*. Princeton, New Jersey: Princeton University Press, 2003.

[10] W. Wagenaar, *Paradoxes of Gambling Behavior*. Mahwah, New Jersey: Lawrence Erlbaum, 1984.

[11] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–292, March 1979.

[12] D. Prelec, "The probability weighting function," *Econometrica*, vol. 4, no. 3, pp. 497–527, May 1998.

[13] R. Gonzales and G. Wu, "On the shape of the probability weighting function," *Cognitive Psychology*, vol. 38, no. 1, pp. 129–166, 1999.

[14] C. Camerer and T. Ho, "Experience-weighted attraction learning in normal form games," *Econometrica*, vol. 26, no. 4, pp. 827–874, 1999.

[15] R. Hertwig, G. Barron, E. U. Weber, and I. Erev, "Decisions from experience and the effect of rare events in risky choice." *Psychological Science*, vol. 15, no. 8, pp. 534 – 539, 2004.