Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis

Ming Cao, Andrew Stewart, and Naomi Ehrich Leonard

***Abstract*— Leveraging research by psychologists on human decision-making, we present a human-robot decision-making problem associated with a complex task and study the corresponding joint decision-making dynamics. The collaborative task is designed so that the human makes decisions just as human subjects make decisions in the two-alternative, forced-choice task, a well-studied decision-making task in behavioral experiments. The human subject chooses between two options at regular time intervals and receives a reward after each choice; for a variety of reward structures, the behavioral experiments show convergence to suboptimal choices. We propose a human-supervised robot foraging problem in which the human supervisor makes a sequence of binary decisions to assign the role of each robot in a group in response to a report from the robots on their resource return. We discuss conditions under which the decision dynamics of this human-robot task is reasonably well approximated by the kinds of reward structures studied in the psychology experiments. Using the Win-Stay, Lose-Switch human decision-making model, we prove convergence to the experimentally observed aggregate human decision-making behavior for reward structures with matching points. Finally, we propose an adaptive law for robot reward feedback designed to help the human make optimal decisions.**

I. Introduction

For highly complex tasks with many elements and many scales, there is an important role for automation, where fast, dedicated data processing and feedback responsiveness can be exploited, see, e.g., cooperative control of multi-agent systems [1]. In many complex tasks, such as air traffic control [2] or missions where the environment changes with time and unanticipated events are frequent, it can also be critically important to keep humans in the loop and engaged in the overall decision-making. This allows to exploit their superior ability to handle the unexpected and to recognize pattern and extract structure from data.

It is thus of great interest to investigate how humans and robots can best *jointly* contribute to decision-making [3]. Research in human-robot interaction [4] makes clear that profitable integration of human and robot decision-making dynamics should take advantage of strengths of human decision-makers and of robotic agents. A major challenge is understanding how humans make decisions and what are their associated strengths and weaknesses.

Our approach is to leverage the experimental and modeling work of psychologists and behavioral scientists on human decision-making. To do this we seek commonality in the kinds of decisions humans make in complex tasks and the kinds of decisions humans make in psychology experiments. We consider a class of sequential binary decision-making called the two-alternative forced-choice task for which there is ample research [5], [6], [7], [8], [9], [10]. In this task, the human subject in the psychology experiments chooses between two options at regular time intervals and receives a reward after each choice that depends on recent past decisions. Interestingly, these experiments show convergence of the aggregate behavior to rewards that are often suboptimal.

To apply the behavioral research we introduce a joint human-robot decision-making task associated with a complex task in which the human's role can be mapped into the two-alternative forced-choice task. We assume an environment where human decision-making is needed, i.e., a fully automated decision-making system could potentially fail. The setting is a human-supervised collective robotic foraging problem, where a group of robots moves around in a highly uncertain field and collects a distributed resource and a human supervisor sequentially assigns the role of each of the robots, either to be an explorer or an exploiter of resource. The human and robots work as a team to maximize resource collected.

In our framework the human decision-making takes the form of a two-alternative forced-choice task where the reward report is a feedback from the robots. We discuss conditions under which the reward structures used in the psychology experiments provide reasonable approximations of the human-robot task so that we can apply the results from the psychology literature to study how the human will behave in the complex task. Using the Win-Stay, Lose-Switch (WSLS) human decision-making model together with a model of the two-alternative forced-choice task, we prove convergence of the human behavior to the observed aggregate decision-making for reward structures with matching points. Since behavior converges to suboptimal performance, we propose an adaptive law for the robot feedback that uses only local information but helps the human make optimal decisions.

We review the two-alternative forced-choice tasks in Section II. In Section III we present a map from the decision-making task of the human supervisor of a robotic foraging team to the two-alternative forced-choice task. In Section IV we present our model. We prove convergence of the model in Section V. In Section VI we present our adaptation law for computational aid to human decision-making.

M. Cao is with Faculty of Mathematics and Natural Sciences, ITM, University of Groningen, the Netherlands (m.cao@rug.nl). A. Stewart and N. E. Leonard are with Department of Mechanical and Aerospace Engineering, Princeton University, USA ({arstewar,naomi}@princeton.edu).

This research was supported in part by AFOSR grant FA9550-07-1-0-0528 and ONR grants N00014–02–1–0826 and N00014–04–1–0534.

II. Binary Decision Making Problems

Real-world decision-making problems are difficult to study since the reward for a decision usually depends in a nontrivial way on the decision history. Many studies have considered decision-reward relationships that are fixed; however, these have limited value in addressing problems associated with complex, time-varying tasks. Here, we briefly review a class of decision-making tasks called the *two-alternative forced-choice task*, where reward depends on past decisions.

Montague et al. [9], [6] introduced a dynamic economic game with a series of decision-reward relationships that depend on a subject's decision history. The human subject is faced with a two-alternative sequential choice task. Choices of either A or B are made sequentially and a reward for each decision is administered directly following the choice. Without knowing the reward structure, the human subject tries to maximize the total reward (sum of sequence of rewards).

Two reward structures considered are the *matching shoulder* (shown in Figure 1) and *rising optimum* (shown in Figure 2). In each of these figures, the reward (f_A) for choosing A and the reward (f_B) for choosing B is plotted as a function of the fraction of times A was chosen in the previous $N = 20$ trials ($y = \#A\text{'s}/20$). For example, in Figure 1, if the subject always chooses A, the reward drops to below 0.2. Subsequently, if B is chosen, the reward jumps up close to 1.0. However, continued choices of B lead to declining reward. The average reward, plotted as a dashed line on each figure, is computed as $yf_A(y) + (1-y)f_B(y)$. The optimal strategy is the one that maximizes the average reward curve.

Herrnstein [5], [11] pointed out that, in experiments, human subjects tend to adopt strategies that bring them close to the *matching point* of the reward curves (where $f_A = f_B$). This is reasonable since near the matching point the reward for choosing A or B is about the same. However, this implies that humans do not necessarily converge on the optimal strategy, since the matching point does not necessarily correspond to the optimal average reward.

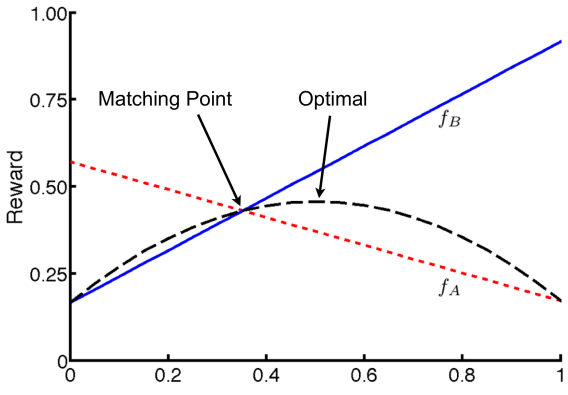

Fig. 1. The *matching shoulder* reward structure [6]. The dotted line depicts f_A, the reward for choice A. The solid line depicts f_B, the reward for choice B. The dashed line is the average value of the reward. Each is plotted against the proportion of choice A made in the last 20 trials.

In Figure 1 the optimal reward is to the right of the matching point. The rising optimum reward structure of Figure 2 is an even more dramatic case. The optimal strategy for the rising optimum case corresponds to choosing option A 100% of the time. The reward at the matching point is significantly suboptimal and to reach the optimum, the human subject must first endure very low rewards.

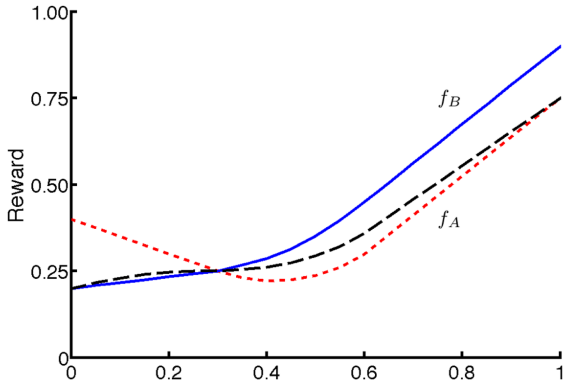

Fig. 2. The *rising optimum* reward structure [10]. Note that, in this case, the value of the reward at the matching point is significantly lower than the optimal reward value.

III. Human-Supervised Robotic Foraging Problem

In this section we formulate a human-supervised robotic foraging problem where the human makes sequential binary decisions and we define an explicit map from the human-supervised robotic foraging problem to a two-alternative forced-choice task. We assume that the environment is changing with time and unanticipated events are frequent so that we need a human in the loop, engaged in the decision-making. Our focus is on integrating decision-making dynamics of the human and the robots; development of collective foraging strategies for the robots is not covered in this paper.

Consider a team of $N = 20$ autonomous robots, foraging in a spatially distributed field $\mathcal{S}$, that are *remotely supervised* by a human. Each robot forages in one of two modes: to collect resources while moving around or to collect resources at its current position. We say that a robot is *exploring* if it is in the former mode or *exploiting* otherwise. The role of the human supervisor is to make the choice for each individual robot, sequentially in time, as to whether it should explore or exploit. At time $t = 0, 1, 2, \ldots$, the human supervisor chooses one of the two modes for a robot: to explore or to exploit. The robot then provides the supervisor with an estimate of the amount of resource to be collected in the next time period under the assigned foraging mode. That estimate represents the reward for the supervisor's decision at time t. By reading robots' reports and making sequential decisions, the human supervisor allocates each of the 20 robots to foraging modes, one at a time, with the objective of maximizing the total resource collection. The human continues to re-assign robots' foraging modes as long as necessary; for example, in a changing environment re-allocation may be critical.

There are alternative possibilities for representation of the reward reported by the robots. For example, the reward could be the total amount of resource collected after some time period. This introduces a time delay in the decision-making process which differs from the fast time-scales in the psychology studies of [7] and [10]. However, the influence of the timing has been studied experimentally and it has been shown that the convergence to matching behavior is actually enhanced with slower time-scales [5].

The role of the human supervisor in the robotic foraging problem is analogous to the role of the human subject in the psychology studies reviewed in Section II. By assumption, however, the human-supervised robot foraging task is more complex than the tasks that the human subjects carry out; notably the dependence of the reward on the past $N = 20$ decisions in the complex task may be more dynamic than reflected in the reward structures shown in Figures 1 and 2. We are exploring conditions under which the task from the psychology studies provides a good enough approximation to the complex task so that we can apply results from the psychology studies and our corresponding models and analysis to understand the human-robot decision dynamics in the complex task. Below we discuss plausible scenarios in the robotic foraging task that map to the the matching shoulder and rising optimum reward structures of Figures 1 and 2. A stronger justification of the map can be found in [12] where we derive a reward curve from numerical experiments of a robot foraging team collecting resource in a simulated environment with multiple high-density patches of resource; the reward curve averaged over a large number of experiments takes the shape of the rising optimum reward structure of Figure 2.

We take $N = 20$ in the rest of the paper since this is a common choice in the psychology experiments. However, the concepts, models and analysis are valid for general N.

A. Matching shoulder

Assume that there are sufficient resources in $\mathcal{S}$ that are uniformly distributed, allowing a robot to collect resources whether exploring or exploiting. We denote the number of robots that are exploring at time t, $t = 0, 1, 2, \ldots$, by $n_A(t)$. When $n_A(t) > 0$ and the mode of a robot i, $1 \leq i \leq 20$, is changed from exploiting to exploring, robot i has to spend some time planning its path to avoid competing with the $n_A(t)$ foraging robots. As a result, the amount of resource expected to be collected during $(t, t+1]$ is less than one and is reasonably well described by

$$1 - w_A(n_A(t) + b_A) \tag{1}$$

where w_A is a weight on the number of exploring robots and b_A is a bias term. The number of exploiting robots at time t is $20 - n_A(t)$. It is assumed that if at time t robot i's mode is changed from exploring to exploiting there is a cost associated with communicating with the $20 - n_A(t)$ exploiting robots to register its own current position and as a result the expected amount of resource to be collected by robot i during $(t, t + 1]$ is reasonably well described by

$$1 - w_B((20 - n_A(t)) + b_B) \tag{2}$$

where w_B is a weight on the number of exploiting robots and b_B is a bias against exploring robots.

Reward reports corresponding to equations (1) and (2) map to matching shoulder reward curves. For example, for $w_A = \frac{1}{50}, b_A = \frac{150}{7}, w_B = \frac{3}{80}$, and $b_B = \frac{20}{9}$, the reward curves are $f_A = -\frac{2}{5}y + \frac{4}{7}$ and $f_B = \frac{3}{4}y + \frac{1}{6}$, plotted in Figure 1.

B. Rising optimum

As shown in Figure 2, the reward curves in the rising optimum task rise quickly and monotonically after the fraction of choice A becomes greater than $\frac{1}{2}$. This reward structure describes reasonably well a robotic foraging situation in which the cost of collaboration outweighs the reward from collaboration up to some threshold in the fraction of explorers, after which cooperation comes with a higher reward. For example, suppose the resources collected by an exploring robot decreases when $n_A(t)$ grows and $n_A(t) \leq 10$. When $n_A > 10$, suppose that exploring robots benefit from collaboration in a way that outweighs the cost of collaborating. In this case the reported reward will map to rising optimum reward curves.

IV. Decision-Making Model

Experimental studies of human subjects performing the two-alternative, forced-choice task with matching shoulder and rising optimum reward structures show consistent aggregate behavior that converges to the allocation of A's that corresponds to the matching point, see, e.g., [9]. Such matching tendency in humans and animals was first identified by Herrnstein [5], [11], whose related work has been influential in quantitative analysis of behavioral and mathematical psychology. However, few mathematically provable results, which describe the tendency for matching behavior, have been obtained and reported. This is, in part, due to the difficulty in modeling the dynamics of human and animal decision-making. Montague and Berns [6] show that the matching point is an attracting point, although their argument requires a limiting assumption. In [13] the authors perform a related analysis.

Several models have been proposed to describe the dynamics of human decision-making. In this paper we analyze the *Win-Stay, Lose-Switch* (WSLS) model, also known as *Win-Stay, Lose-Shift*, which is used in psychology, game theory, statistics and machine learning [14], [15].

Let $x_1(t) \in \{A, B\}$ denote the decision for the binary choice A or B at time t and let $x_i(t) = x_1(t - i + 1)$, $i = 2, \ldots, 20$, denote the decisions of the finite past. Here, the "sliding window" of the last 20 trials is big enough, as validated by experimental data, to include all the past choices that may affect the subject's current decision. Let y denote the fraction of choice A in the last 20 trials, and thus

$$y(t) = \frac{1}{20} \sum_{i=1}^{20} \delta_{iA}(t) \tag{3}$$

where

$$\delta_{iA}(t) = \begin{cases} 1 & \text{if } x_i(t) = A \\ 0 & \text{if } x_i(t) = B. \end{cases} \quad (4)$$

Note that y can only take value from a finite set $\mathcal{Y}$ of twenty-one discrete values:

$$\mathcal{Y} = \{jc,\ j = 0, 1, \ldots, 20\} \text{ where } c = \frac{1}{N} = \frac{1}{20}.$$

The reward at time t is given by

$$r(t) = \begin{cases} f_A(y(t)) & \text{if } x_1(t) = A \\ f_B(y(t)) & \text{if } x_1(t) = B\,. \end{cases} \quad (5)$$

The matching shoulder decision-reward relationship is described by

$$f_A = k_A y + c_A \quad (6)$$

$$f_B = k_B y + c_B \quad (7)$$

where k_A and k_B are the slopes and c_A and c_B are the constant terms of the two given linear reward curves.

From the definitions of x_i, $2 \leq i \leq 20$, we also have

$$x_i(t+1) = x_{i-1}(t), \qquad i = 2, \ldots, 20,\ t = 0, 1, 2, \ldots \quad (8)$$

Thus, the human decision-making process in the matching shoulder case can be modeled as a twenty-dimensional, discrete-time dynamical system described by equations (3)-(8) where $x_i(t)$, $1 \leq i \leq 20$, is the state of the system and $y(t)$ is the output of the system.

The human subject's decisions may be affected by all the decisions and rewards in the past trials. In fact, a goal of the studies of two-alternative forced-choice tasks is to determine the decision-making mechanism through experiment and behavioral and neurobiological investigations. Here we consider the WSLS model. This model assumes that human decisions are made with information from the rewards of the previous two choices only and that a switch in choice is made when a decrease in reward is experienced. To summarize,

$$x_1(t+1) = \begin{cases} x_1(t) & \text{if } r(t) \geq r(t-1); \\ \bar{x}_1(t) & \text{otherwise,} \end{cases} \quad t = 1, 2, 3, \ldots \quad (9)$$

where $\bar{\cdot}$ denotes the "not" operator; i.e. if $x_1(t) = A$ (resp. $x_1(t) = B$), then $\bar{x}_1(t) = B$ (resp. $\bar{x}_1(t) = A$).

This model is especially interesting in the setting of human-supervised robotic foraging tasks, which has been discussed in Section III, because a similar decision rule is used in [16] to explain the ability of foraging predators to converge to the *ideal free distribution*, the optimal allocation of density of individuals to territories with various resource levels [17]. Individuals visit a territory and immediately determine the profitability, q, specific to that area. Profitability of a territory is assumed to decrease monotonically with respect to density. If q is greater than, or equal to, the environmental average γ, individuals remain in that territory. Otherwise, they leave and visit another territory. Individuals must, however, learn the environmental average as they visit territories using the following algorithm:

$$\gamma(t+1) = \alpha q(t) + (1-\alpha)\gamma(t) \quad (10)$$

where α is a constant in the interval $[0, 1]$. When α is equal to 1 the decision rule is equivalent to WSLS.

In the sequel, we give a rigorous analysis of the dynamics of human performance in tasks with matching shoulder rewards. It is shown that for the human decision-making model (9), the fraction of choice A converges to a neighborhood of the matching point. We note that the result applies locally to the rising optimum curves, which have the same structure as the matching shoulder curves in a neighborhood of the matching point.

V. Convergence Analysis

In this section, we analyze the convergence behavior of the system (3)-(9). We assume that the matching shoulder reward curves (6) and (7) satisfy

$$k_A < 0,\ k_B > 0, \text{ and } f_A,\ f_B \text{ intersect.} \quad (11)$$

Let y^* denote the value of y at the matching point, namely the intersection of the two lines (6) and (7). We consider the general case when

$$y^* \notin \mathcal{Y}. \quad (12)$$

As shown in [12], limit cycles can occur when $y^* < \frac{1}{3}$ or $y^* > \frac{2}{3}$ for certain reward structures with the WSLS model. This behavior, however, has not been observed in the decision-making studies. Nonetheless, to formally rule out these limit cycles with the WSLS model, we require that the reward curves f_A and f_B satisfy

$$\frac{1}{3} \leq y^* \leq \frac{2}{3}. \quad (13)$$

The linear curves used in the experiments [6] satisfy the conditions (11), (12) and (13), so the analysis in this section provides an analytical understanding of human decision-making dynamics in two-alternative forced-choice tasks of the same type.

A. Convergence of WSLS

The following result describes the oscillating behavior of $y(t)$ near y^*. Let y^l denote the greatest element in $\mathcal{Y}$ that is smaller than y^* and let y^u denote the smallest element in $\mathcal{Y}$ that is greater than y^*.

Theorem 1: For system (3)-(9) satisfying conditions (11)-(13), if $y(t_1) \in \mathcal{L} = [y^l, y^u]$ for some $t_1 > 0$, then $y(t) \in \mathcal{L}' = [y^l - c, y^u + c]$ for all $t \geq t_1$.

Remark 1: In this paper we consider $N = 20$ robots for accurate correspondence with [10]. This can be generalized and the convergence result applies for arbitrary $N \geq 6$. We present a proof of this, and a discussion of the effect of larger or smaller N on convergence rate, in [12].

Remark 2: Condition (11) can be relaxed to include nonlinear f_A and f_B which satisfy the property that f_A and f_B decrease monotonically with fraction of choice A and choice B, respectively. In [12] we prove convergence for this case.

Remark 3: Some nonlinear reward structures, such as the rising optimum of Figure 2, have local regions with the structure of Remark 2. In such examples, convergence to a region containing the matching point applies locally.

Remark 4: Condition (12) is not necessary for convergence. In fact, in the case that $y^* \in \mathcal{Y}$, a tighter convergence result applies. This result is presented in [12].

To prove Theorem 1, we need the following four lemmas, the proofs of which are presented in [12].

Lemma 1: For system (3)-(9), with conditions (11)-(13) satisfied, if $x_1(t_1) = A$, $x_1(t_1+1) = A$ and $y(t_1) < 1$ for some $t_1 \geq 0$, then there exists $0 \leq \tau \leq 20$ such that $y(t) = y(t_1)$ for $t_1 \leq t \leq t_1+\tau$ and $y(t_1+\tau+1) = y(t_1)+c$.

Lemma 2: For system (3)-(9), with conditions (11)-(13) satisfied, if $x_1(t_1) = B$, $x_1(t_1+1) = B$ and $y(t_1) > 0$ for some $t_1 \geq 0$, then there exists $0 \leq \tau \leq 20$ such that $y(t) = y(t_1)$ for $t_1 \leq t \leq t_1+\tau$ and $y(t_1+\tau+1) = y(t_1)-c$.

Lemma 3: For system (3)-(9), with conditions (11)-(13) satisfied, if $y(t_1) < y^*$ and $y(t_1+1) = y(t_1) - c > 0$ for some $t_1 \geq 0$, then there exists $0 \leq \tau \leq 20$ such that

$$y(t) = y(t_1) - c \ \text{ for } t_1 \leq t \leq t_1 + \tau \tag{14}$$

and

$$y(t_1 + \tau + 1) = y(t_1). \tag{15}$$

Lemma 4: For system (3)-(9), with conditions (11)-(13) satisfied, if $y(t_1) > y^*$ and $y(t_1+1) = y(t_1) + c$ for some $t_1 \geq 0$, then there exists $0 \leq \tau \leq 20$ such that

$$y(t) = y(t_1) + c \ \text{ for } t_1 \leq t \leq t_1 + \tau \tag{16}$$

and

$$y(t_1 + \tau + 1) = y(t_1). \tag{17}$$

Now we are in a position to prove Theorem 1.

Proof of Theorem 1: If $y(t) \in \mathcal{L}$ for all $t \geq t_1$, then the conclusion holds trivially. Now suppose this is not true. Let $t_2 > t_1$ be the first time for which $y(t) \notin \mathcal{L}$. Then it suffices to prove the claim that the trajectory of $y(t)$ starting at $y(t_2)$ stays at $y(t_2)$ for a finite time and then enters $\mathcal{L}$. Note that $y(t_2)$ equals either $y^l - c$ or $y^u + c$. Suppose $y(t_2) = y^l - c$, then the claim follows directly from Lemma 3; if on the other hand, $y(t_2) = y^u + c$, then the claim follows directly from Lemma 4. □

Theorem 1 gives the convergence analysis in the neighborhood $\mathcal{L}$ of the matching point y^*. Our next step is to present the global convergence analysis for the system (3)-(8). It is easy to check that if the system starts with the initial condition $y(0) = 0$ and $x_1(1) = B$ or the initial condition $y(0) = 1$ and $x_1(1) = A$, then the trajectory of $y(t)$ will stay at its initial location. In what follows, we show that if the trajectory of $y(t)$ starts in $(0, 1)$ and conditions (11)-(13) are satisfied, then the trajectory always enters $\mathcal{L}$ after a finite time.

Proposition 1: For any initial condition of the system (3)-(9) satisfying $0 < y(0) < 1$ and suppose conditions (11)-(13) are satisfied, there is a finite time $T > 0$ such that $y(T) \in \mathcal{L}$.

To prove Proposition 1, we need the following four lemmas, the proofs of which are also contained in [12].

Lemma 5: For system (3)-(9), with conditions (11)-(13) satisfied, if $y(t_1) < y^*$, $y(t_1+1) = y(t_1)$ and $x_1(t_1+1) \neq x_1(t_1)$ for some $t_1 \geq 0$, then there exists a finite $\tau > 0$ such that

$$y(t_1 + \tau) = y(t_1) + c. \tag{18}$$

Lemma 6: For system (3)-(9), with conditions (11)-(13) satisfied, if $y(t_1) > y^*$, $y(t_1+1) = y(t_1)$ and $x_1(t_1+1) \neq x_1(t_1)$ for some $t_1 \geq 0$, then there exists a finite $\tau > 0$ such that

$$y(t_1 + \tau) = y(t_1) - c. \tag{19}$$

Lemma 7: For system (3)-(9), with conditions (11)-(13) satisfied, if $0 < y(t_1) < y^l$ and $y(t_1+1) = y(t_1) - c$ for some $t_1 \geq 0$, then there exists a finite $\tau > 0$ such that

$$y(t_1 + \tau) = y(t_1) + c. \tag{20}$$

Lemma 8: For system (3)-(9), with conditions (11)-(13) satisfied, if $y^u < y(t_1) < 1$ and $y(t_1+1) = y(t_1) + c$ for some $t_1 \geq 0$, then there exists a finite $\tau > 0$ such that

$$y(t_1 + \tau) = y(t_1) - c. \tag{21}$$

Now we are in a position to prove Proposition 1.

Proof of Proposition 1: For any $0 < y(0) < 1$, either $y(1) = y(0) + c$, or $y(1) = y(0)$, or $y(1) = y(0) - c$. We will discuss these three possibilities in each of two cases. First consider the case where $y(0) < y^l$. If $y(1) = y(0) - c$, according to Lemma 7, there is a finite time t_1 for which $y(t_1) > y(0)$. If $y(1) = y(0)$ and $x_1(1) \neq x_1(0)$, according to Lemma 5, there is a finite time t_2 for which $y(t_2) > y(0)$. If $y(1) = y(0)$ and $x(1) = x(0) = A$, according to Lemma 1, there is a finite time t_3 for which $y(t_3) > y(0)$. If $y(1) = y(0)$ and $x(1) = x(0) = B$, according to Lemma 2, there is a finite time $\bar{t}_4$ for which $y(\bar{t}_4 - 1) = y(0)$ and $y(\bar{t}_4) = y(\bar{t}_4 - 1) - c$. Then according to Lemma 7, there is a finite time t_4 for which $y(t_4) > y(0)$. So for all possibilities of $y(1)$ there is always a finite time $\bar{t} \in \{1, t_1, t_2, t_3, t_4\}$ for which $y(\bar{t}) > y(0)$. Using this argument repeatedly, we know that there exists a finite time T_1 at which $y(T_1) = y^l \in \mathcal{L}$. Now consider the other case where $y(0) > y^u$, then using similar arguments, one can check that there exists a finite time T_2 for which $y(T_2) = y^u \in \mathcal{L}$. Hence, we have proven the existence of T which lies in the set $\{T_1, T_2\}$. □

Combining the conclusions in Theorem 1 and Proposition 1, we have proven the following theorem, which describes the global convergence property of $y(t)$.

Theorem 2: For any initial condition of the system (3)-(9) satisfying $0 < y(0) < 1$ and suppose conditions (11)-(13) are satisfied, there exists a finite time $T > 0$ such that for any $t \geq T$, $y(t) \in \mathcal{L}'$.

In the next section, we discuss how the local and global convergence properties of matching behavior can be utilized to improve performance of the robotic system with a human decision-maker introduced as in Section III.

VI. Adaptive Reward Feedback

In this section we propose a means to improve decision-making performance using the integrated human-robot team by taking advantage of the tendency for humans to converge to the matching point in a class of reward curves. We consider the human-supervised robot foraging problem in the case that the robot reported reward has a convergent matching point that does not coincide with the optimum (e.g. Figure 1). Our approach is to have the robots manipulate their reported reward, i.e., have them adapt their feedback, to aid

the human in finding the optimal strategy. This adaptive law requires only that the robots keep track of recent supervisor decisions and corresponding rewards.

As proved in Section V, for a class of reward curves, given a generic initial condition, a human decision-maker will converge to a region $\mathcal{L}'$ which contains y^*, the value of y at the matching point. By recording decisions made and rewards reported in $\mathcal{L}$ and $\mathcal{L}'$, the robots can make a local approximation of the reward curves f_A and f_B and, in particular, estimate the gradient of the average reward defined as $g(y) = \frac{d}{dy}\left(y\, f_A(y) + (1-y) f_B(y)\right)$. Adaptive feedback is used if $g(y^*) \neq 0$ and there does not exist $y_{opt} \in \mathcal{L}$ such that $g(y_{opt}) = 0$, as this would mean that $\mathcal{L}$ contains a value of y which maximizes the average reward. For example, in the case of reward curves shown in Figure 1, y^* and y_{opt} differ by more than $3c$ and the human decision-maker will typically converge to a region $\mathcal{L}'$ that does not contain y_{opt}.

In such a scenario, we propose to have the robots use adapted curves, $\bar{f}_A$ and $\bar{f}_B$, to determine the reward. Specifically, we vary $\bar{f}_A$ and $\bar{f}_B$ so that the matching point moves in the direction of y_{opt}. In this paper we present a strategy that varies the slopes k_A and k_B. To ensure stability and avoid confusing the supervisor, adapted curves are chosen so that the new y^* changes by no more than $c = 1/N$ each time an adaptation occurs. We denote the reward curves of the q^{th} adaptation by $\bar{f}_{A,q}$ and $\bar{f}_{B,q}$ with $\bar{f}_{A,0} = f_A$ and $\bar{f}_{B,0} = f_B$; the value of y at the matching point is y^*_q. The slopes of $\bar{f}_{A,q+1}$ and $\bar{f}_{B,q+1}$ ($\bar{k}_{A,q+1}$ and $\bar{k}_{B,q+1}$) are updated by increasing $\bar{k}_{A,q+1}$ and decreasing $\bar{k}_{B,q+1}$ both by dk, where

$$dk = \frac{1}{2}\left(\frac{(k_A - k_B)(c_B - c_A)}{c(k_A - k_B) + c_B - c_A} + k_B - k_A\right).$$

This implies that $y^*_{q+1} = y^*_q + c$. We use $-dk$ in place of dk to achieve $y^*_{q+1} = y^*_q - c$.

Let $\mathcal{L}_q = [y^l_q, y^u_q]$ where $y^l_q = \max\{y \in \mathcal{Y} | y < y^*_q\}$ and $y^l_q = \min\{y \in \mathcal{Y} | y > y^*_q\}$. For rapid convergence to the new y^*_{q+1}, curves $f_{A,q+1}$ and $f_{B,q+1}$ are introduced only when one of the following conditions is met:

$$y(t) = y^u_q \quad \text{and} \quad g(y^*_q) > 0 \tag{22}$$

$$y(t) = y^l_q \quad \text{and} \quad g(y^*_q) < 0. \tag{23}$$

The human decision-maker will subsequently converge to the matching point of the adapted curves while the matching point of the adapted curves approaches y_{opt}. In [12] we prove an upper bound on the time required to converge to y^*_{q+1} once $\bar{f}_{A,q+1}$ and $\bar{f}_{B,q+1}$ are introduced.

VII. Concluding Remarks

A natural extension to the framework developed in this paper is to consider N clusters of M robots each, where the reward curves will likely depend on M. Of similar interest is to consider dynamically changing N. The latter will change the resolution of the reward feedback and may serve as a useful strategy in aiding a decision-maker to find the optimum. In ongoing work, we are pursuing similar analyses with other models that have been used successfully to describe human behavior in two-alternative forced-choice tasks for a range of decision-reward relationships. This framework will be useful to consider multiple human decision-makers in concert with the results discussed in [10].

VIII. Acknowledgement

We thank Jonathan Cohen, Damon Tomlin and Patrick Simen from Princeton's Center for the Study of Brain, Mind and Behavior for discussions on experiments and modeling of the two-alternative forced-choice tasks. Likewise, we thank Phil Holmes and Andrea Nedic for their feedback on modeling and Debbie Prentice for social psychology input.

References

[1] P. Antsaklis and J. Baillieul, editors. *Proc. IEEE: Special Issue on Technology of Networked Control Systems*, volume 95:1. IEEE, 2007.

[2] C. Niessen and K. Eyferth. A model of the air traffic controller's picture. *Safety Science*, 37:187–202, 2001.

[3] J.L. Burke, R.R. Murphy, E. Rogers, V.J. Lumelksy, and J. Scholtz. Final report for the DARPA/NSF interdisciplinary study on human-robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34:103–112, 2004.

[4] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 35(4):460–470, 2005.

[5] R. Herrnstein. Rational choice theory: necessary but not sufficient. *American Psychologist*, 45:356–367, 1990.

[6] P. R. Montague and G. S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36:265–284, 2002.

[7] 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*, 1153:111–121, 2007.

[8] J. Li, S. M. McClure, B. King-Casas, and P. R. Montague. Policy adjustment in a dynamic economic game. *PLoS One*, e103:1–11, 2006.

[9] D. M. Egelman, C. Person, and P. R. Montague. A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience*, 10:623–630, 1998.

[10] A. Nedic, D. Tomlin, P. Holmes, D.A. Prentice, and J.D. Cohen. A simple decision task in a social context: preliminary experiments and a model. In *Proc. 47th IEEE Conf. Decision and Control*, 2008.

[11] R. Herrnstein. *The Matching Law: Papers in Psychology and Economics*. Harvard University Press, Cambridge, MA, USA, 1997. Edited by Howard Rachlin and David I. Laibson.

[12] M. Cao, A. Stewart, and N.E. Leonard. Convergence in human decision-making dynamics and integration of humans with robots. http://www.princeton.edu/~naomi/publications.html, 2008.

[13] L. Vu and K. Morgansen. Modeling and analysis of dynamic decision making in sequential two-choice tasks. In *Proc. 47th IEEE Conf. Decision and Control*, 2008.

[14] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of American Mathematical Society*, 58:527–535, 1952.

[15] M. Nowak and K. Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, 364:56–58, 1993.

[16] C. Bernstein, A. Kacelnik, and J.R. Krebs. Individual decisions and the distribution of predators in a patchy environment. *Journal of Animal Ecology*, 57:1007–1026, 1988.

[17] S.D. Fretwell and H.L. Lucas. On territorial behavior and other factors influencing habitat distribution in birds. *Acta Bio.*, 19:16–36, 1970.