INVITED PAPER

Towards Human–Robot Teams: Model-Based Analysis of Human Decision Making in Two-Alternative Choice Tasks With Social Feedback

A Markov model of decision making in human groups is used to develop predictive capability; analysis yields testable results on sensitivity of individual performance.

By Andrew Stewart, *Member IEEE*, Ming Cao, *Member IEEE*, Andrea Nedic, *Student Member IEEE*, Damon Tomlin, and Naomi Ehrich Leonard, *Fellow IEEE*

ABSTRACT | With a principled methodology for systematic design of human–robot decision-making teams as a motivating goal, we seek an analytic, model-based description of the influence of team and network design parameters on decision-making performance. Given that there are few reliably predictive models of human decision making, we consider the relatively well-understood two-alternative choice tasks from cognitive psychology, where individuals make sequential decisions with limited information, and we study a stochastic decision-making model, which has been successfully fitted to human behavioral and neural data for a range of such tasks. We use an extension of the model, fitted to experimental data from groups of humans performing the same task simultaneously and receiving feedback on the choices of others in the group. First, we show how the task and model can be regarded as a Markov process. Then, we derive analytically the steady-state probability distributions for decisions and performance as a function of model and design parameters such as the strength and path of the social feedback. Finally, we discuss application to human–robot team and network design and next steps with a multirobot testbed.

KEYWORDS | Decision making; human machine systems; multiagent systems; psychology

Manuscript received October 3, 2010; revised August 4, 2011; accepted October 5, 2011. Date of publication December 13, 2011; date of current version February 17, 2012. This work was supported in part by the U.S. Air Force Office of Scientific Research (AFOSR) under Grant FA9550-07-1-0-0528.

A. Stewart was with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 USA. He is now with the Applied Physics Laboratory, University of Washington, Seattle, WA 98105-6698 USA (e-mail: andy@apl.washington.edu).
M. Cao is with the Faculty of Mathematics and Natural Sciences, Institute of Technology, Engineering and Management (ITM), University of Groningen, 9700 AK Groningen, The Netherlands (e-mail: m.cao@rug.nl).
A. Nedic was with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA. She is now with Princeton Consultants, Princeton, NJ 08540 USA (e-mail: anedic@princeton.com).
D. Tomlin is with the Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544 USA (e-mail: nedic@princeton.edu; dtomlin@princeton.edu).
N. Ehrich Leonard is with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: naomi@princeton.edu).

Digital Object Identifier: 10.1109/JPROC.2011.2173815

I. INTRODUCTION

There is growing interest in enabling humans and robots to jointly make decisions that address problems in a variety of complex tasks, such as information gathering in an uncertain, dynamic environment [1], search and rescue [2], and characterization of a hazardous environment [3], where neither a fully automated nor a fully manual operation is sufficient. This motivates development of a principled

methodology that systematizes the design of human–robot decision-making teams. Such a methodology should leverage strengths and compensate for weaknesses in both the humans and the robots who constitute the team.

In [4]–[6], it is argued that for humans and robots to best leverage each other's strengths, they should collaborate as peers. This means expanding from the more traditional view of humans as supervisors of robots to one in which robots can make their own decisions. In these works, there is an emphasis on adjustable autonomy, i.e., robots work autonomously but query humans as needed. In [4], humans and robots work in parallel on a task, resolving problems through interaction. Conditions for when and how humans and robots should communicate are presented in [6].

We too are interested in designing decision-making teams of humans and robots acting as peers; however, our approach is to make use of developments in cognitive psychology and rigorously examine human decision-making dynamics with social interactions. In this paper, we seek to derive analytic expressions for decision-making performance of human–robot teams as a function of design parameters. These include those that define the social context, such as the size of the team, the distribution of different types of decision makers on the team, and the social feedback network, that is, how information flows among human and robot team members. Such expressions make it possible to systematically select design parameters so that decision-making teams meet performance requirements.

The general problem of rigorously examining social decision making presents a major challenge because the space of decision-making problems is vast, and there are few rigorously derived and analytically tractable models that reliably predict how humans make decisions. Accordingly, in this paper, we focus on a relatively well-understood family of tasks from the cognitive psychology literature and a well-tested model for human decision making that has recently been extended and fitted to experimental data for groups of human decision makers in a social context.

This family of tasks, known as two-alternative forced-choice (TAFC) tasks, has been used in human decision-making experiments and studies to investigate a variety of fundamental questions, such as how humans trade off exploration versus exploitation to find optimal decision strategies [7]–[9] and how humans often settle for suboptimal strategies [10]–[12]. A TAFC task requires a human subject to make a sequence of choices between two known alternatives. After every choice, the subject receives a score that serves as a reward, and the subject's goal is to maximize accumulated reward over the entire sequence of choices.

By manipulating the reward structure, the task can be changed to represent different kinds of decision-making challenges. We consider four prototypical TAFC tasks from the literature, each corresponding to a different reward structure as described in Section II. In each of the four tasks, the reward depends not only on the most recent choice but also on a recent finite history of choices. In every case, there is an optimal sequence of choices, i.e., one that maximizes average reward; however, the amount of exploration required to find the optimal sequence varies among the four tasks.

We base our analysis on a stochastic soft-max choice model used in the cognitive psychology literature to predict how a human makes decisions in TAFC tasks. The successful fitting of both human behavioral and neural [functional magnetic resonance imaging (fMRI)] data taken during TAFC task experiments [7] justifies the use of the model to describe a human decision maker in TAFC tasks. It is shown in [13] that the soft-max choice model emerges from a drift-diffusion (DD) equation. For empirical work that justifies using the DD model in perceptual TAFC tasks, see, for example, [14]–[16]. Derivations and applications of the DD equation for decision making are treated comprehensively in [13]. There it is shown that the DD model is the continuum limit of the sequential probability ratio test for binary hypothesis testing from statistical decision theory [17], [18]. The DD model can also be derived from the dynamics of a variable that represents the evidence in neuronal populations in favor of one alternative over the other [13].

Adopting the soft-max choice model for our analysis is further motivated by the recent extension and empirical fitting of this model to decision making in groups of humans performing the TAFC tasks of Section II [8], [19]. In the experiments of [8] and [19], human subjects, working in parallel on the same TAFC task, made simultaneous choices while receiving social feedback. The social feedback was provided after each choice: each human subject received not only his or her own reward, but also a report on the current choice and/or reward of the other subjects. The goal of these experiments was to explore the role of this kind of limited group information feedback in individual decision making in the TAFC task setting. The extended model uses a soft-max choice model for each decision maker and couples the multiple models with behaviors representing individual responses to the choices and/or rewards of others.

In this paper, we focus on choice feedback among groups of decision makers engaged in parallel in the same TAFC task. Our central contribution is to use the extended model of [8], which has already been fitted to experimental data, to derive *analytic* predictions of the decisions and performance as a function of model parameters. With a small number of reasonable, simplifying assumptions, we make the problem tractable by showing it can be represented as a low-dimensional Markov process. We derive probability distributions of steady-state decision sequences and corresponding steady-state performance as a function of parameters that characterize the task, the individual

decision makers, and the feedback network. We show that our analytic predictions produce the same trends as those produced with the original empirically validated model of [8] and also those from the experiments.

Using our analytic predictions, we examine how performance varies as a function of model parameters, and we consider how to choose those that are design parameters to enforce decision-making performance as desired. For example, we can choose how many and which human team members to include. If we model each robotic decision maker on the team with a soft-max choice model, then we can design the characterizing parameters for each robotic team member. We can also design the choice feedback network, i.e., who gets information from whom. We describe new experiments underway that will test predicted performance of designed human–robot systems.

Because of the generalizability of TAFC tasks, our analysis has potential applicability to real-world problems beyond those that map well into binary decision-making tasks. As in many real-world problems, the TAFC tasks require a decision maker to evaluate information from the environment and from other decision makers and to respond with a decision. While the output of a single binary decision is rather simple, the evaluative process required to optimally perform a TAFC task is complex. A decision maker has to make do with a limited measure of the environment that takes value in an unknown set and changes with every decision. In the case with choice feedback, a decision maker can also evaluate the choices of others, which take value in a binary set, but which may appear to be in conflict with the information associated with measurements from the environment. The TAFC tasks we study represent a range of explore-versus-exploit challenges, and the learning and implementation of explore-versus-exploit policies are pervasive components of any decision-making process in an unknown environment.

The paper is organized as follows. The TAFC tasks are described in Section II. The model for decision making is described in Section III. We analyze decision making and derive steady-state probability distributions as a function of model and task parameters in Section IV for a decision maker alone, and in Section V, for a group of decision makers in the social context. We conclude and discuss application to human–robot team design and next steps with a multirobot testbed in Section VI. Proofs can be found in the Appendix. Preliminary results have appeared in [20] and [21].

II. TWO-ALTERNATIVE FORCED-CHOICE TASK

A. Task Description

In the TAFC task introduced in [7], [9], a human subject is prompted by a computer to choose between two alternatives (denoted A and B) within a fixed period of time after the prompt. Once a choice is made and the "A" or "B" button is pushed, the computer reports a score that represents a reward (performance measure), and the task repeats. The subject's goal is to maximize total accumulated reward over the duration of the task (optimize performance over the long run); at the end of the experiment, the subject is paid in proportion to the sum of rewards received.

In [7], each experiment consisted of 250 sequential decisions. In [8] and [19], each experiment consisted of 150 sequential decisions with a fixed period of 1.7 s for response after the prompt; if the subject failed to enter a choice within the allotted time, the system recorded the same choice as was made at the last decision time.

Subjects are not told that their reward depends on their recent choice history. The number of immediate past choices N fixes the extent of choice history that determines the reward. The choice history $y(t)$ at time t is the proportion of choices of A in the most recent N choices. Let $i(t)$ be the number of times A was chosen in the most recent N choices up to time t, then $y(t) := i(t)/N$. Note that y belongs to a finite, discrete set given by $\mathcal{Y} = \{(i/N), i = 0, 1, \ldots, N\}$. In the experiments of [8], [19], and [22], $N = 20$, and in [7], $N = 40$; these values push the limits of what a human subject can remember.

Fig. 1 shows the four reward structures that we examine in this paper; these reward structures are all defined and used in the literature; see, e.g., [7]–[9], [19], and [22]. Each reward structure in Fig. 1 is defined by two curves of reward as a function of y: the dashed red curve plots $r_A(y)$, the reward received in the case that button A is pushed, and the solid blue curve plots $r_B(y)$, the reward received in the case that button B is pushed. The long-dashed black curve plots the average value of reward $\bar{r}(y) = yr_A(y) + (1-y)r_B(y)$.

A choice sequence y is optimal if it maximizes $\bar{r}(y)$. A local maximum can be found by climbing the gradient in $\bar{r}$; however, this is difficult in practice. Each of the four reward structures in Fig. 1 defines a different task with its own decision-making challenge; the four tasks range in how much the human subject must explore to find an optimal choice sequence. We refer to a task as "easy" if the optimum can be found with little exploration and "difficult" if it is only found with considerable exploration. The value of y that maximizes $\bar{r}(y)$ is not necessarily an element in $\mathcal{Y}$, but there is always a $y \in \mathcal{Y}$ that differs from the optimal value by less than $1/N$.

1) Matching Shoulders Task: The matching shoulders (MS) reward structure is illustrated in Fig. 1(a). For the MS task, the reward curves are lines: $r_A(y) = k_A y + c_A$ and $r_B(y) = k_B y + c_B$. In Fig. 1(a), $k_A = -0.5$, $c_A = 0.6$, $k_B = 1$, and $c_B = 0$. So, for example, if the decision maker

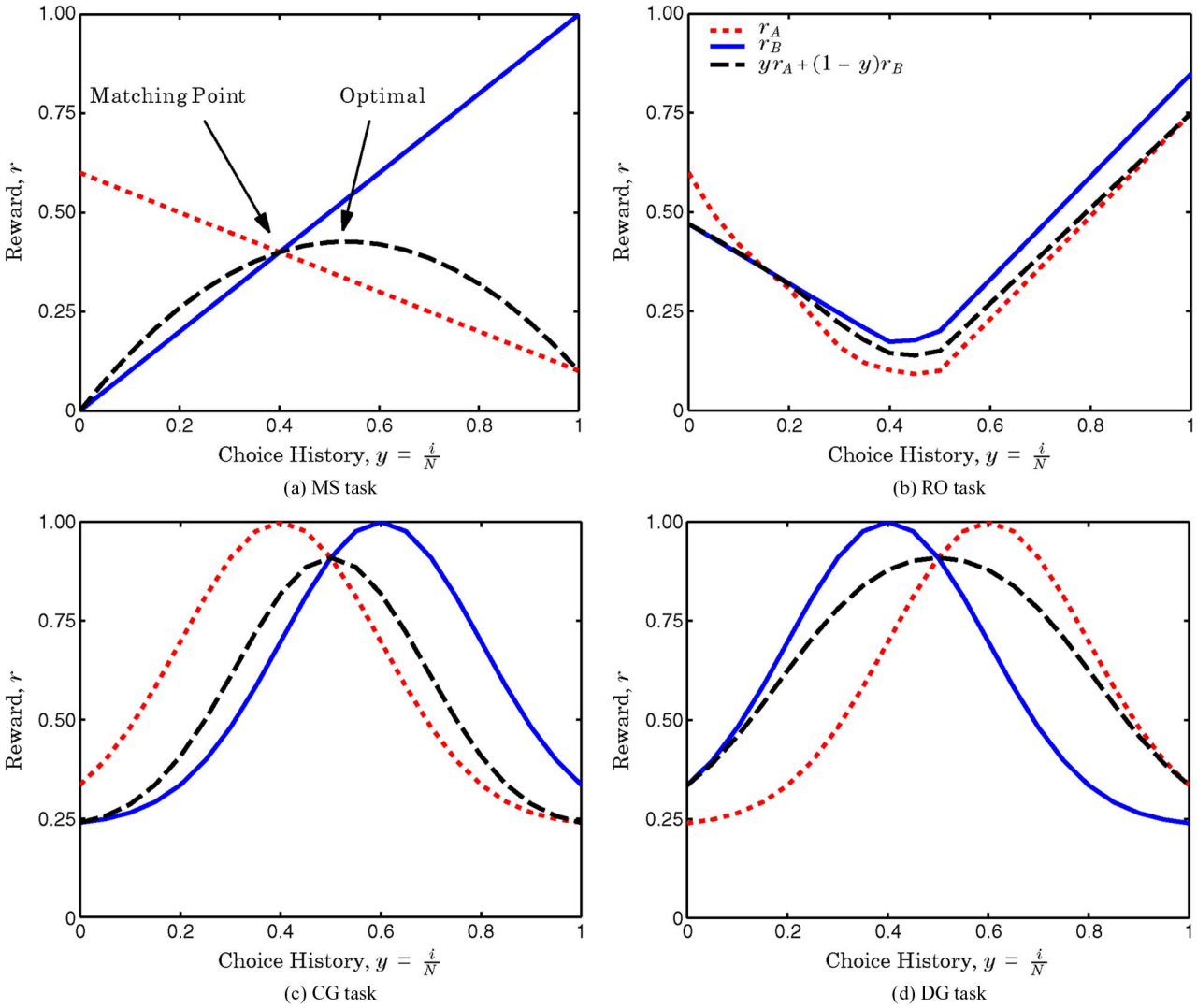

Fig. 1. ***Four reward structures: (a) matching shoulders (MS); (b) rising optimum (RO); (c) converging Gaussians (CG); and (d) diverging Gaussians (DG). In each plot, the dashed red curve is r_A, the reward for choice A, and the solid blue curve is r_B, the reward for choice B. The long-dashed black curve is the average value of the reward $\bar{r}$. Each is plotted against choice history $y = i/N$, $i = 0, 1, 2, \ldots, N$, i.e., the proportion of choice A in the last N choices.***

has chosen A half of the time in the last N trials of the task, then a current choice of A yields a reward of $r_A(0.5) = 0.35$ and a current choice of B yields a reward of $r_B(0.5) = 0.5$. The value of y that maximizes $\bar{r}$ is found by solving $(d/dy)\bar{r}(y) = 0$ to get $y = (k_B + c_A - c_B)/(2(k_B - k_A))$. In Fig. 1(a), $\bar{r}(y)$ has a unique maximum at $y = 0.53$.

The slopes of the MS reward curves represent two resources A and B that have diminishing returns. Indeed the reward for choosing A drops the more frequently A is chosen and likewise the reward for choosing B drops the more frequently B is chosen. An optimal choice sequence corresponds to $y = 0.53$; however, it is challenging for the subject to find this optimal value. This is because the point at which the two curves intersect $(y = 0.4)$, called the *matching point*, is an attractor.

To see this, note from Fig. 1(a) that the decision maker receives a higher reward for choosing B rather than A whenever $y > 0.4$. However, continued choice of B reduces y and when $y < 0.4$, the decision maker will find that choosing A yields a higher reward than choosing B. Subsequent choices of A will increase y and the process repeats once $y > 0.4$ again. Indeed, there is extensive empirical evidence that human decision makers converge in aggregate to choice sequences y that correspond to the matching point [7]–[9]. The consequence of matching behavior leading to suboptimal choices was studied extensively by Herrnstein [10]–[12]. Conditions for convergence of human decision making to the matching point have been proved and analyzed using decision-making models in [9] and [23]–[25].

In order to formalize the role of social feedback in tasks with an attracting matching point, such as the MS task, we first use our analysis in Section IV to examine the influence of model parameters for individual decision makers on matching and suboptimal decision making.

2) Rising Optimum Task: The rising optimum (RO) reward structure of Fig. 1(b) also has a matching point, but it is a more complex task since there is a local optimum at $y = 0$ and a global optimum at $y = 1$. This RO reward structure is studied with and without social feedback in [8] with subjects who begin the task with the initial condition $y(0) = 0$. Subjects tend to spend time at the local optimum or near the matching point, but rarely find the global optimum, since to do so requires making choice sequences along the way that yield the lowest possible rewards in the task [in Fig. 1(b), these are choice sequences in the range $y = 0.4$ to $y = 0.5$].

Even if a subject reaches the optimum at $y = 1$, a choice of B will yield an even higher reward than a choice of A—this will reduce y, moving the decision maker away from the optimal solution. Thus, considerable exploration is needed to find the optimal choice sequence, and so the RO task is a difficult task. An important question is how to design human–robot teams with the right kind of feedback so that they perform better in a difficult task like the RO task as compared to individuals who do not share information with one another. We use our analytic predictions to systematically explore the implications of feedback in the RO task in Section V-D.

3) Converging and Diverging Gaussians Tasks: Fig. 1(c) and (d) shows converging Gaussians (CG) and diverging Gaussians (DG) reward structures, respectively. These two structures differ only in that what is r_A in the CG structure is r_B in the DG structure and what is r_B in the CG structure is r_A in the DG structure. The implication of the difference is inherent in the name of the tasks. In the CG task, the matching point is an attractor such that decision makers tend to converge to it, whereas in the DG task the matching point is divergent such that decision makers tend to move away from it. Both structures are symmetric about $y = 0.5$, which corresponds both to the matching point and to the optimal decision-making solution.

The CG task is an easy task since the matching point, and therefore the optimal solution, is attracting, and thus very little exploration is needed to find it. The DG task is more difficult since the optimal solution is divergent. The DG task was designed in [8] to enable exploratory behavior to split decision makers into arbitrary groups on either side of the symmetry point $y = 0.5$, and thus allowing the impact of social feedback to be investigated.

B. Task Model

Let $x(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ denote the last N choices of the decision maker ordered sequentially in time with $x_1(t) \in \{A, B\}$ denoting the decision at time t, $x_2(t) \in \{A, B\}$ the decision at time $t-1$, etc., i.e.,

$$x_k(t+1) = x_{k-1}(t), \qquad k = 2, \ldots, N \\ t = 0, 1, 2, \ldots. \tag{1}$$

The choice history, i.e., the proportion of choice A in the last N decisions at time t, can be computed from $x(t)$ as

$$y(t) = \frac{1}{N} \sum_{k=1}^{N} \delta_{kA}(t) \tag{2}$$

where $\delta_{kA}(t) = 1$ if $x_k(t) = A$ and $\delta_{kA}(t) = 0$ if $x_k(t) = B$. The reward $r(t)$ at time t is given by

$$r(t) = \begin{cases} r_A(y(t)), & \text{if } x_1(t) = A \\ r_B(y(t)), & \text{if } x_1(t) = B. \end{cases} \tag{3}$$

We define the reward difference as

$$\Delta r(y(t)) := r_B(y(t)) - r_A(y(t)). \tag{4}$$

The variables $x(t)$ and $y(t)$ evolve according to a stochastic decision-making process, presented in Section III, and thus are treated as random variables.

C. Task With Social Feedback

The RO, CG, and DG tasks were all used in the experiments with and without social feedback as described in [8] and [19]. In each experiment, five human subjects, physically isolated from one another, made choices in parallel for the same task at the same time. In experiments without social feedback, the human subjects were in the "alone condition." In experiments with social feedback, after every choice when the computer reported the reward, it also reported the current choice of each of the other four subjects (choice feedback), the current reward of each of the other four subjects (reward feedback), or both the current choice and reward of the other four subjects (choice and reward feedback). Accordingly, in the experiments with social feedback, each human subject could use the information reported about the four others in their own decision making. In this paper, we restrict our analysis of social feedback to the case of choice feedback.

In the experiments of [8] and [19], the feedback on choices and/or rewards passed from every individual to every other individual in the group, i.e., interconnections were undirected and the network graph was complete. These experiments were intentionally designed as a first

step to investigate social feedback in the case that each subject's role was the same and information was shared equally.

It is of great interest to understand how social feedback affects decision-making performance in the case of more general feedback interconnections. In this paper, we study networks of human decision makers with choice feedback paths defined by directed graphs as well as the undirected graph used in the experiments.

In the directed case, we study the decision-making dynamics of a focal individual who receives feedback on the choices of M other decision makers, each of whom receives no social feedback. We make a formal comparison between the influence of directed versus undirected choice feedback in Section V-E, and we describe new human subject experiments underway that test our predictions in the case of directed choice feedback in Section V-D. The directed case allows us to investigate the influence of designed (i.e., robotic) peer decision makers who provide feedback but do not change their own strategies; we discuss additional experiments underway in this context.

III. DECISION-MAKING MODEL

A. Soft-Max Choice Model

The stochastic soft-max choice model was first proposed by Egelman *et al.* [7] to describe human decision making in TAFC tasks of the kind presented in Section II. This model saw continued use by Montague and Berns [9] and has since become a well-accepted decision-making model in this context. The model prescribes the probability $p_A(t+1) := \Pr\{x_1(t+1) = A\}$ that a subject will choose A at time $t+1$ as a sigmoidal function of the state at time t

$$p_A(t+1) = \frac{1}{1+e^{-\mu(w_A(t)-w_B(t))}}. \tag{5}$$

The probability p_A depends explicitly on the difference between the subject's *anticipated* reward w_A for choosing A next and the subject's *anticipated* reward w_B for choosing B next. Note that w_A and w_B are independent from r_A and r_B, i.e., w_A and w_B are modeled as having been determined by the subject according to a learning process, described below.

Fig. 2 shows p_A as a function of $w_A - w_B$ in the case that $\mu = 1$. The parameter μ determines the slope of the sigmoidal function. Larger μ implies more certainty in decision making, which can be interpreted as less of a tendency to explore. As μ tends to infinity, (5) becomes deterministic: in this case, whenever $w_A > w_B (w_A < w_B)$, a choice of $A(B)$ is made. In [24], we proved convergence results for this deterministic limit.

As shown in [8], the choice model best predicts a subject's choice sequences in the RO task when $\mu = 11.0$

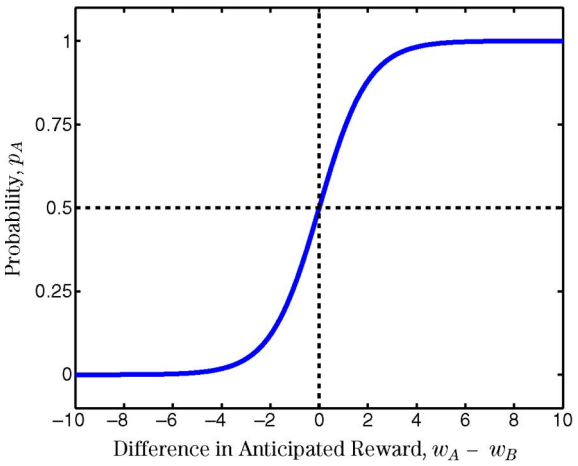

Fig. 2. *Sigmoidal function given by (5) representing probability p_A of choosing A as a function of $w_A - w_B$ (plotted here with $\mu = 1$).*

and in the CG task when $\mu = 2.5$. These values correspond to fits over all subjects, but fits of μ were also made to individual subjects. It is a goal of this work to develop a formal understanding of the effects on performance of parameters such as μ. In the social context, a heterogeneous group of decision makers can be studied by distinguishing individuals by their own characteristic value of μ.

Inspired by temporal difference learning, the studies in [7] of the role of dopamine neurons in coding for reward prediction error [26] have suggested a discrete-time linear model of the dynamic update of w_A and w_B. Let $Z \in \{A, B\}$ be the choice made at time t, then

$$w_Z(t+1) = (1-\lambda)w_Z(t) + \lambda r(t) \tag{6}$$

$$w_{\bar{Z}}(t+1) = w_{\bar{Z}}(t), \qquad t = 0, 1, 2, \ldots \tag{7}$$

where $\bar{Z}$ denotes the alternative choice to Z. Here, $\lambda \in [0, 1]$ is a learning rate. Larger λ implies less "memory"; when $\lambda = 1$ there is no memory since the anticipated reward is equal to the most recent reward received.

The model (5) has the same form as that predicted by the stochastic differential equation that describes a scalar DD process used widely to model perceptual decision making [13], [27], [28]

$$dz = \alpha dt + \sigma dW, \qquad z(0) = 0. \tag{8}$$

Here z represents the accumulated evidence in favor of a candidate choice of interest (e.g., choice A), α is a drift rate representing the signal intensity of the stimulus acting on z, and σdW is a Wiener process with standard deviation σ, which is the diffusion rate representing the effect of white noise. On each trial a choice of A is made when $z(t)$ first

crosses the predetermined threshold $+\xi$ and a choice of B is made when $z(t)$ first crosses $-\xi$. It can be computed using tools developed in [13] that the probability of choosing A in the next time step is given by (5) with the appropriate mapping between parameters.

B. Soft-Max Choice Model With Social Feedback

Each individual in a group of decision makers can be modeled with a soft-max choice model as described above. In [8] and [19], social feedback was introduced with a feedback term that interconnects the individual choice models. Models with different numbers of fitting parameters were compared using the Akaike information criterion together with estimated maximum likelihoods for the prediction of choice sequences. The following choice feedback model performed well in those tests.

Consider a focal decision maker who receives choice feedback from M other decision makers. The favored choice feedback model of [8] and [19] biases the focal decision maker's anticipated rewards with a feedback parameter ν that reinforces his/her tendency to choose $A(B)$ when a majority of the M others chooses $A(B)$. The probability that the focal individual chooses A is

$$p_A(t+1,\nu) = \frac{1}{1+e^{-\mu(w_A(t)-w_B(t)+\nu u(t))}} \tag{9}$$

$$u(t) = \begin{cases} 1, & \text{if } |A| \geq \left\lceil \dfrac{M+1}{2} \right\rceil \\ -1, & \text{if } |B| \geq \left\lceil \dfrac{M+1}{2} \right\rceil \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $|A|$ is the number of others who choose A at time t (similarly for $|B|$), and $\lceil \cdot \rceil$ gives the smallest integer greater than or equal to its argument. The no-feedback case (5) is equivalent to $p_A(t+1,0)$ in (9).

C. Assumptions for Analysis

The state of the choice model for a single decision maker in the TAFC task is the N-element decision history $x(t)$ and the two anticipated rewards $w_A(t)$ and $w_B(t)$. In this section, we define two assumptions that reduce the state to the scalar choice history $y(t)$; we make these assumptions in our analysis for the remainder of the paper. Since $x(t)$ evolves according to a stochastic process defined by the model and $y(t)$ is computed from $x(t)$, we treat $y(t)$ as a random variable in our analysis.

We make Assumption 1 for all reward structures. We make Assumption 2a for the MS, CG, and DG reward structures and Assumption 2b for the RO reward structure. Recall that $\Delta r(y) := r_B(y) - r_A(y)$.

Assumption 1: $Pr\{x_k(t) = A | x(t)\} = y(t)$.

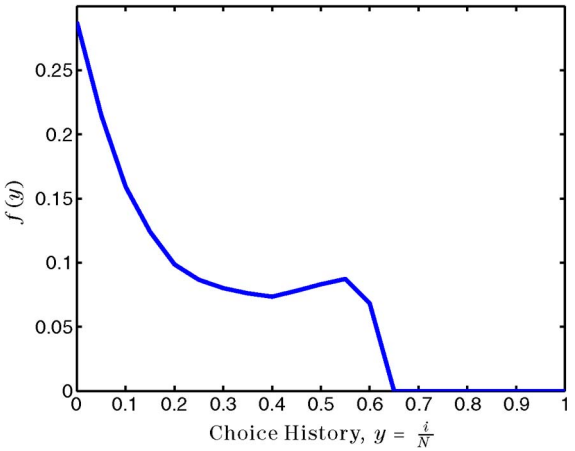

Fig. 3. *Computed difference in anticipated reward $f(y)$ used in Assumption 2b for the RO reward structure (shown for $N = 20$).*

Assumption 2a: $w_B(t) - w_A(t) = \Delta r(y(t))$.

Assumption 2b: $w_B(t) - w_A(t) = f(y(t))$, where $f(y)$ is given by the curve in Fig. 3.

Assumption 1 implies that the yN A's and $(1-y)N$ B's in $x(t)$ are uniformly distributed in the choice history. When Assumption 1 holds, the state of the system can be represented by $y(t)$, $w_A(t)$, and $w_B(t)$. Assumption 1 is believed to hold when the decision making occurs over long time periods [9], since, for each $y(t)$ visited by the system, all possible combinations of ordering of choices within $x(t)$ should occur with approximately equal frequency. Extensive numerical simulations of the soft-max choice model support the assumption.

Assumption 2a sets the difference in the subject's anticipated rewards at time t equal to the difference in actual rewards evaluated at $y(t)$. The subject knows the actual reward at $y(t)$ corresponding to his/her choice at time t, and may have recently learned the actual reward for the other alternative at $y(t)$. When Assumptions 1 and 2a hold, the system state can be represented by $y(t)$.

Assumption 2a was introduced by Montague and Berns [9] in their analysis of attraction to the matching point in reward functions. To further investigate we performed a numerical study without using Assumption 2a by building a Markov chain with state $y(t)$, $w_A(t)$, and $w_B(t)$. We used $\lambda = 1$ as this approximates well the fitted value of λ in the CG and DG tasks [8]. We computed an equilibrium distribution for this case and observed that it varied insignificantly from the equilibrium distribution derived in the case where we let Assumption 2a hold (see Section IV-B). This result supports Assumption 2a.

Assumption 2b is specific to the RO reward structure of Fig. 1(b) where Assumption 2a does not apply. The function $f(y)$, shown in Fig. 3, is determined from a simulation in which the model (5)–(7) made choice sequences in the

RO task and the computed anticipated rewards for each value of y were averaged. In the simulation, $\lambda = 0.1$, which approximates well the fitted value of λ for the RO task [8]. In this difficult RO task, the subject does not make choice A often enough to achieve $y(t) \geq 0.65$, and so $f(y) = 0$ for $y \geq 0.65$.

IV. ANALYSIS OF DECISION MAKING IN THE ALONE CONDITION

In this section, we analyze a single model decision maker performing the TAFC tasks of Section II. In Section IV-A, we show how to model the system as a Markov process. In Section IV-B, we derive analytically the steady-state probability distribution for that process, i.e., the long-run probability that a decision maker will make choice sequences corresponding to $y \in \mathcal{Y}$. We compare the derived distributions to distributions computed from experimental data. We then use the derived distributions to study performance. We prove conditions under which the decision maker will converge to a matching point in Section IV-C; since matching behavior is critical to how much exploration is needed to make optimal choice sequences in TAFC tasks, this is an important result both for individual and social decision making. In Section IV-D, we derive sensitivity of performance to the model parameter μ, which quantifies an individual's tendency to explore.

A. Markov Model

The transition probabilities for the Markov model of the system with random variable $y \in \mathcal{Y}$ as its state can be computed as follows.

Proposition 1: Suppose Assumptions 1 and 2a hold. Then, the choice model (5) for the TAFC task (1)–(3) is a Markov process with state $y(t)$ and transition probabilities given by

$$\Pr\left\{y(t+1) = y(t) - \frac{1}{N}\right\} = \frac{e^{\mu\Delta r} y(t)}{1+e^{\mu\Delta r}} \tag{11}$$

$$\Pr\{y(t+1) = y(t)\} = \frac{e^{\mu\Delta r} + (1-e^{\mu\Delta r})y(t)}{1+e^{\mu\Delta r}} \tag{12}$$

$$\Pr\left\{y(t+1) = y(t) + \frac{1}{N}\right\} = \frac{1-y(t)}{1+e^{\mu\Delta r}} \tag{13}$$

where $\Delta r = \Delta r(y(t))$ is given by (4). In case Assumption 2b holds instead of Assumption 2a, then the transition probabilities are given by (11)–(13) with $\Delta r(y(t))$ replaced with $f(y(t))$.

Equations (11)–(13) are used to build the $(N+1) \times (N+1)$ state transition matrix $\mathbf{P}$ which has entries $P_{ij} = \Pr\{y(t+1) = (j/N) | y(t) = (i/N)\}$, $i, j \in \{0, 1, \ldots, N\}$.

B. Steady-State Choice Distribution

The state transition matrix $\mathbf{P}$ is tridiagonal, and all tridiagonal elements are positive. So, any state can be reached from any another in finite time, guaranteeing irreducibility. It is aperiodic since return to state i from state i can happen as quickly as one time step, but no state is absorbing. Thus, the process has a unique limiting distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_N)$ describing the fraction of time the process will spend in each of the enumerated states $y = i/N$, $i = 0, 1, \ldots, N$, in the long run (as $t \to \infty$) [29]. This steady-state distribution satisfies

$$\pi \mathbf{P} = \pi \tag{14}$$

$$\sum_{i=0}^{N} \pi_i = 1. \tag{15}$$

Proposition 2: For the transition probabilities given by (11)–(13), the unique steady-state distribution is

$$\pi_i = \frac{\alpha_i \left(1 + e^{\mu\Delta r\left(\frac{i}{N}\right)}\right) e^{-\mu\beta_i}}{\sum_{j=0}^{N} \alpha_j e^{-\mu\beta_j} \left(1 + e^{\mu\Delta r\left(\frac{j}{N}\right)}\right)} \tag{16}$$

where $\alpha_i = N!/((N-i)!i!)$ and $\beta_i = \sum_{j=1}^{i} \Delta r(j/N)$.

The distribution π from (16) is plotted in Fig. 4 for each of the four reward structures of Fig. 1. Fig. 4(a) shows that the decision maker in the MS task primarily makes choices that keep y near the matching point ($y = 0.4$) rather than near the optimal solution at the peak of the average reward curve ($y = 0.53$). Fig. 4(b) shows that the decision maker in the difficult RO task is unable to find the global optimum at $y = 1$ as observed in the experiments of [8]. Instead, time is spent at the local optimum ($y = 0$) and near the matching point. Fig. 4(c) shows for the easy CG task that the decision maker spends most time at the optimum ($y = 0.5$), coincident with the matching point. As shown in Fig. 4(d), the DG task is more difficult than the CG task since decision makers are attracted to either side away from the optimum ($y = 0.5$).

The distribution of Fig. 4(a) for the MS task agrees with experimental results in the literature, e.g., see [7, Fig. 2]. In Fig. 5, we plot the distributions of Fig. 4(b)–4(d) and compare to the distributions from the experiments of the RO, CG, and DG tasks reported in [8]. These experimental distributions are computed as the percentage of time spent with each possible choice history $y \in \mathcal{Y}$, where $N = 20$, averaged over all subjects in the alone condition.

The comparison in Fig. 5(a) shows that the prediction captures the difficulty for a decision maker to find the optimal solution in the difficult RO task and the relatively long time that the decision maker will spend near the local

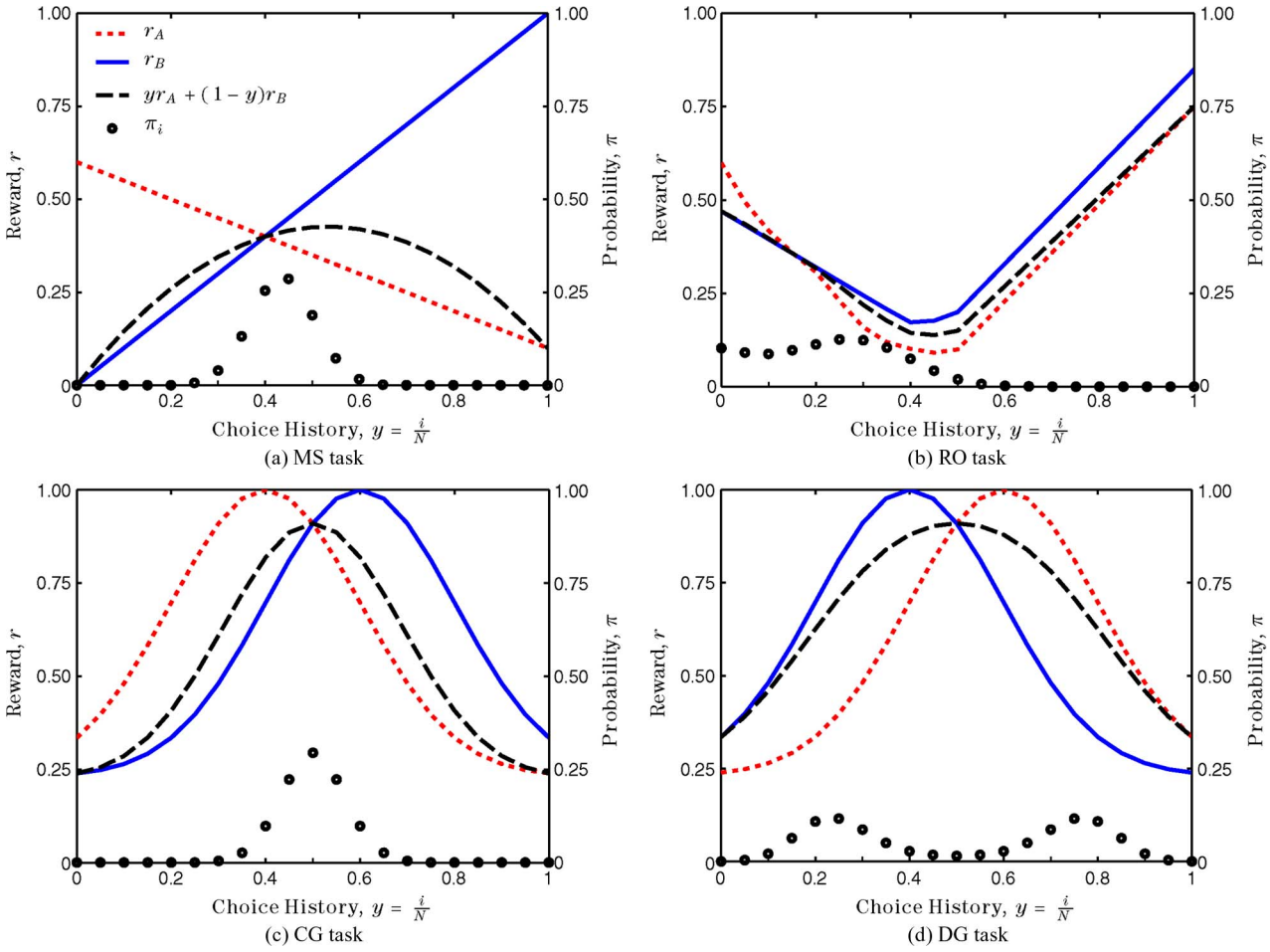

Fig. 4. *Steady-state probability distribution π of y from (16) for TAFC tasks with $N = 20$: (a) MS with $\mu = 5$; (b) RO with $\mu = 11$; (c) CG with $\mu = 2.5$; and (d) DG with $\mu = 2.91$. The probability π_i describes the time the decision maker spends making choice sequences corresponding to $y = i/N$; π_i is plotted with the symbol "o" for each $i = 0, 1, \ldots, N$ for all tasks. The dashed red curve is r_A, the solid blue curve is r_B, and the long-dashed black curve is $\bar{r}$. The values of μ for the RO, CG, and DG tasks are from the best fit to experimental data [8].*

optimum ($y = 0$) and the matching point. The prediction misses the very few human subjects who did find the optimal solution in the experiments. The prediction also slightly overpredicts time spent near the matching point and slightly underpredicts time spent near the local optimum. These differences may be attributable to error introduced in the estimation of $f(y)$ of Assumption 2b.

The comparison in Fig. 5(b) for the CG task is excellent, while the comparison in Fig. 5(c) shows some differences. We expect that these differences are due to the convergence rate of decision making, which is slower in the more challenging DG task as compared to the CG task. Each experiment in [8] consisted of 150 sequential choices. It is likely that this was sufficient in the easy CG task for subjects to converge to a steady-state sequence. However, in the DG task, the subjects do considerable exploring, switching from values of y lower and higher than 0.5, and it may take a longer time for them to settle on one side or the other.

C. Performance and Steady-State Matching

As discussed in Section II, matching behavior, prevalent in TAFC tasks, can have significant implications for how much exploration is needed to find an optimal choice sequence. Near the matching point, the decision maker receives approximately the same reward for choosing A as for choosing B; however, this can be suboptimal in the long run as has been studied extensively by Herrnstein [10]–[12], [30] and more recently by [7]–[9]. Rigorous proofs of matching behavior are limited. In [9], Assumption 2a is used to show that the matching point in the MS reward structure is an attractor. Conditions for convergence to matching are rigorously proved in [24] for the deterministic limit of the choice model (5)–(7). Convergence to matching is proved in [23] and [24] for the win–stay–lose–switch (WSLS) decision-making model. A related analysis for the WSLS model is shown in [25].

In this section, we prove steady-state matching behavior for the choice model (5) by finding sufficient

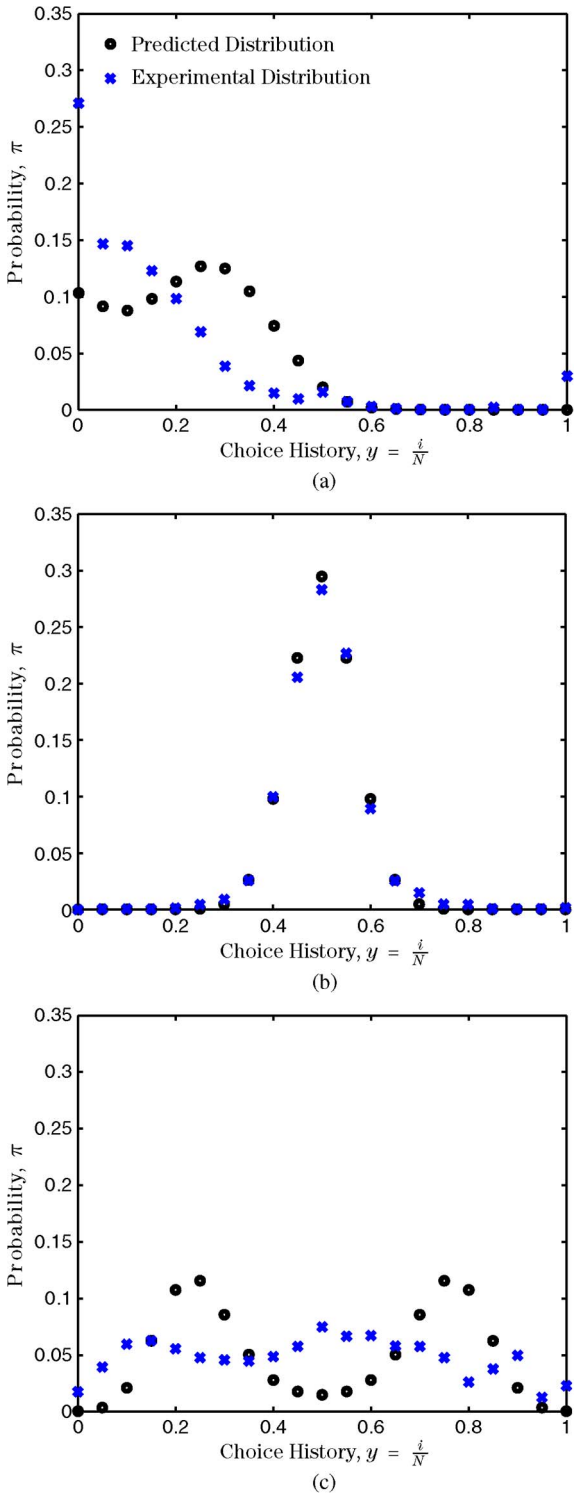

Fig. 5. ***Comparison of steady-state distribution π of y from (16), plotted as black "o," and distributions computed from experimental data from [8], plotted as blue "x." (a) RO task. (b) CG task. (c) DG task.***

conditions on the slope μ that guarantee that π_i is greatest for $y = i/N$ at or near the matching point. In Theorem 1, we find a bound μ_1 such that if $\mu > \mu_1$, then π_i peaks in a small neighborhood of the matching point. In Theorem 2, we find a bound $\mu_2 > \mu_1$ such that if $\mu > \mu_2$, then π_i peaks at the matching point.

Definition 1: A reward structure with a unique matching point of type 1 consists of reward curves $r_A(y)$, $r_B(y)$, for which there exists $y^* = i^*/N$, $i^* \in \{1, 2, \ldots, N-1\}$ that satisfies $\Delta r(y^*) = 0$, $\Delta r(y) < 0$ for $y < y^*$, and $\Delta r(y) > 0$ for $y > y^*$.

Let $\lfloor \cdot \rfloor (\lceil \cdot \rceil)$ be the largest (smallest) integer less than its argument. Define $\gamma = ((N - i^*)!i^*!)/(2\lfloor N/2 \rfloor ! \lceil N/2 \rceil !)$.

Theorem 1: Consider a reward structure with a unique matching point of type 1 and suppose that Assumptions 1 and 2a hold. If

$$\mu > \mu_1 := \max\left\{ \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*+2}{N}\right)}, \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*-2}{N}\right)} \right\} \quad (17)$$

then the steady-state choice distribution is maximum for $y \in \{y^* - (1/N), y^*, y^* + (1/N)\}$.

Theorem 2: Consider a reward structure with a unique matching point of type 1 and suppose that Assumptions 1 and 2a hold. If

$$\mu > \mu_2 := \max\left\{ \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*+1}{N}\right)}, \frac{1-\gamma}{\gamma \Delta r\left(\frac{i^*-1}{N}\right)} \right\} \quad (18)$$

then the steady-state choice distribution is maximum for $y = y^*$.

Example 1: For the MS task of Fig. 1(a), we have $r_A(y) = k_A y + c_A$ and $r_B(y) = k_B y + c_B$, where $k_A = -0.5$, $c_A = 0.6$, $k_B = 1$, and $c_B = 0$. For $N = 20$, by Theorems 1 and 2, $\mu_1 = 5.45$ and $\mu_2 = 10.91$. These values shrink for smaller N and grow for larger N.

Example 2: For the CG task of Fig. 1(c), we have

$$r_A(y) = e^{-\left(\frac{y-\bar{y}_A}{\sqrt{2}\sigma_A}\right)^2} + c_A, \quad r_B(y) = e^{-\left(\frac{y-\bar{y}_B}{\sqrt{2}\sigma_B}\right)^2} + c_B \quad (19)$$

with $\bar{y}_A = 0.4$, $\bar{y}_B = 0.6$ and $\sigma_A = \sigma_B = 0.2$ and $c_A = c_B = 0.3$. For $N = 20$, by Theorems 1 and 2, $\mu_1 = 3.30$ and $\mu_2 = 6.06$.

D. Performance Sensitivity to Model Parameters

Given the analytic expression (16) for π, we compute sensitivity of long-run decision-making performance to the parameter μ. Larger μ corresponds to increased certainty in the decision making, which can also be interpreted as a reduced tendency to explore.

Let $\bar{r}_i$ denote the reward received on average if the decision maker were to maintain choice sequences corresponding $y = i/N$, i.e., $\bar{r}_i := (i/N)r_A(i/N) + (1-(i/N))r_B(i/N)$. Then, the expected value of the reward $\tilde{r}$ is

$$\tilde{r} = \sum_{i=0}^{N} \pi_i \bar{r}_i. \tag{20}$$

The sensitivity of performance to μ is the derivative of the expected value of the reward with respect to μ

$$\frac{d}{d\mu}\tilde{r} = \sum_{i=0}^{N} \bar{r}_i \frac{d}{d\mu}\pi_i = \sum_{i=0}^{N} \left(\frac{i}{N} r_A\left(\frac{i}{N}\right) + \frac{N-i}{N} r_B\left(\frac{i}{N}\right)\right)\frac{d}{d\mu}\pi_i. \tag{21}$$

Denoting $g_i(\mu) := (1 + e^{\mu \Delta r(i/N)})$ and $M(\mu) := \sum_{j=0}^{N} \pi_j$, the derivative of π_i with respect to μ can be written as (22), shown at the bottom of the page.

Example 1 (Continued): Consider again the MS task of Fig. 1(a). The expected value of reward $\tilde{r}$ and its sensitivity to μ given by $(d/d\mu)\tilde{r}$ of (21) are both plotted in Fig. 6 for $N = 20$. The sensitivity has a critical point at $\mu_c = 1.15$. For $\mu < \mu_c$ increasing μ results in substantially higher reward. However, as μ increases further, the expected value of reward decreases. Our analysis precisely describes how much exploratory behavior in the decision making is beneficial. This result is directly related to Theorems 1 and 2: for $\mu > \mu_1 = 5.11$, i.e., with too much certainty (equivalently not enough exploration), the decision maker converges to the matching point, which is not the optimal strategy.

Example 2 (Continued): Consider again the CG task of Fig. 1(c). The expected value of reward $\tilde{r}$ and its sensitivity to μ given by $(d/d\mu)\tilde{r}$ of (21) are both plotted in Fig. 7 for $N = 20$. In this case, $(d/d\mu)\tilde{r}$ is positive for all μ. This is true for any N because the matching point coincides with the maximum of the expected value of reward, and, therefore, increasing the parameter μ (equivalently decreasing exploration in the decision making) results in higher expected reward for the task. We note, however, that there is not a great deal of gain in performance once μ increases above a threshold approximately equal to 5.

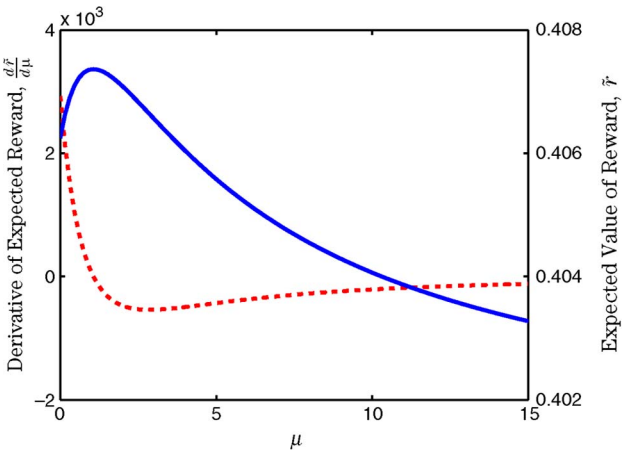

Fig. 6. *Expected value of reward $\tilde{r}$ and sensitivity $(d/d\mu)\tilde{r}$ from (21) for the MS task of Fig. 1(a) for $N = 20$. The solid blue curve is $\tilde{r}$ and the dashed red curve is $(d/d\mu)\tilde{r}$, both plotted as a function of μ.*

V. ANALYSIS OF DECISION MAKING IN A SOCIAL CONTEXT

In this section, we analyze a group of model decision makers who make simultaneous decisions with choice feedback in the TAFC tasks of Section II. We focus first on directed feedback, and we examine the decision dynamics of the focal individual who receives choice feedback from M others, who themselves do not receive any social feedback. In Section V-A, we show how to model the system as an inhomogeneous Markov process. We investigate convergence in Section V-B and derive analytically the expected equilibrium probability distribution for the process. Using this probability distribution in Section V-C, we study the role of choice feedback in performance in the CG task, we compare to experimental data, and we compute sensitivity of performance to team and network parameters μ, ν, and M. In Section V-D, we study the role of choice feedback in performance in the RO task, and we design a heterogeneous team predicted to

$$\frac{d}{d\mu}\pi_i = \frac{\alpha_i e^{-\mu\beta_i}\left(\Delta r\left(\frac{i}{N}\right)e^{\mu\Delta r\left(\frac{i}{N}\right)} - \beta_i g_i(\mu)\right)}{M(\mu)} - \frac{\alpha_i e^{-\mu\beta_i} g_i(\mu)\sum_{j=0}^{N}\alpha_j e^{-\mu\beta_j}\left(\Delta r\left(\frac{j}{N}\right)e^{\mu\Delta r\left(\frac{j}{N}\right)} - \beta_j g_j(\mu)\right)}{M(\mu)^2} \tag{22}$$

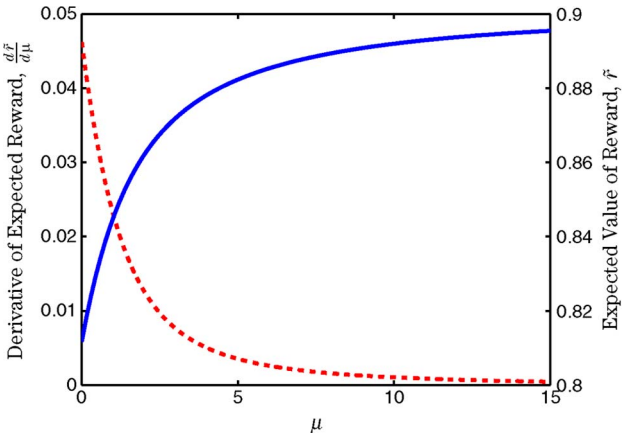

Fig. 7. *Expected value of reward $\tilde{r}$ and sensitivity $(d/d\mu)\tilde{r}$ from (21) for the CG task of Fig. 1(c) for $N = 20$. The solid blue curve is $\tilde{r}$ and the dashed red curve is $(d/d\mu)\tilde{r}$, both plotted as a function of μ.*

improve performance for the focal decision maker. In Section V-E, we analyze the case of undirected choice feedback and compare results to the directed feedback case for the CG and DG tasks.

A. Expectation of the Markov Model

The identification of the dynamics of the focal decision maker with directed choice feedback from M others as a Markov process with state $y(t)$ is analogous to that for the decision maker in the alone condition studied in Section IV-A. However, in the social context, the Markov process is inhomogeneous because at each time t, the one-step state transition matrix depends on the choices of others through the function $u(t)$ in (10). By conditioning on the value of $u(t)$ at each time, we can analyze the expectation of the inhomogeneous process; we derive the expectation of the state transition matrix as follows.

Proposition 3: Suppose Assumptions 1 and 2a hold. Then, the choice model representing the focal individual receiving choice feedback from M others (9)–(10) for the TAFC task (1)–(3) is a Markov process with state $y(t)$ and expected state transition probabilities given by

$$\Pr\left\{y(t+1) = y(t) - \frac{1}{N}\right\} = [1 - \bar{p}_A(y(t))]y(t) \tag{23}$$

$$\Pr\{y(t+1) = y(t)\} = [1 - \bar{p}_A(y(t))] \times (1 - y(t)) + \bar{p}_A(y(t))y(t) \tag{24}$$

$$\Pr\left\{y(t+1) = y(t) + \frac{1}{N}\right\} = \bar{p}_A(y(t))(1 - y(t)) \tag{25}$$

where $\Delta r = \Delta r(y(t))$ is given by (4) and

$$\bar{p}_A(y(t), \nu) = \frac{\Pr\{u(t) = 1\}}{1 + e^{\mu(\Delta r - \nu)}} + \frac{\Pr\{u(t) = -1\}}{1 + e^{\mu(\Delta r + \nu)}} + \frac{\Pr\{u(t) = 0\}}{1 + e^{\mu \Delta r}}. \tag{26}$$

The conditional probabilities on $u(t)$ are given by

$$\Pr\{u(t)=1\} = \sum_{k=\left\lceil \frac{M+1}{2} \right\rceil}^{M} \binom{M}{k} p_A(\infty, 0)^k (1 - p_A(\infty, 0))^{M-k}$$

$$\Pr\{u(t)=-1\} = \sum_{k=\left\lceil \frac{M+1}{2} \right\rceil}^{M} \binom{M}{k} (1 - p_A(\infty, 0))^k p_A(\infty, 0)^{M-k}$$

$$\Pr\{u(t) = 0\} = 1 - (\Pr\{u(t) = 1\} + \Pr\{u(t) = -1\}) \tag{27}$$

with $\binom{M}{k} = M!/(k!(M-k)!)$. In case Assumption 2b holds instead of Assumption 2a, then the results hold with $\Delta r(y(t))$ replaced with $f(y(t))$.

The $(N+1) \times (N+1)$ one-step state transition matrix $P(t)$ has entries

$$P_{ij} = \Pr\left\{y(t+1) = \frac{j}{N} \,|\, y(t) = \frac{i}{N}\right\} \tag{28}$$

$i, j \in \{0, 1, \ldots, N\}$. Using (23)–(25), we can build the expectation $\mathbf{P}$ of this transition matrix.

B. Convergence and Steady-State Choice Distribution

1) Convergence in Probability: In Section V-A, before conditioning on $u(t)$, we have obtained an explicit expression for the $(N+1)$-dimensional state transition matrix $P(t)$. The sequence of $P(t)$ in t is a sequence of independent identically distributed (i.i.d.) matrix-valued random variables. It is easy to check that each $P(t)$ is a tridiagonal matrix in which all the elements on the diagonal, subdiagonal, and superdiagonal are positive. To show that the distribution of y converges in probability, it suffices to show that the matrix $R(t) = P(0)P(1)\cdots P(t)$ converges almost surely to a rank-one matrix $\mathbf{1}\pi$ for some probability distribution π, where $\mathbf{1}$ is the vector of all ones. Towards this end, we introduce the following result from [31].

Proposition 4: The matrix $R(t)$ converges almost surely to a rank-one matrix as t goes to infinity if and only if for

every $i, j \in \{1, \ldots, N+1\}$

$$\Pr\{\mathcal{E}_{ij}\} = 1$$

where

$$\mathcal{E}_{ij} = \left\{\exists k, \exists t | R_{ik}(t) R_{jk}(t) > 0\right\}.$$

This result can be used to show that the distribution of $y(t)$ converges in probability.

Theorem 3: There exists a probability distribution π such that

$$\lim_{t\to\infty} R(t) = \mathbf{1}\pi \qquad \text{almost surely.}$$

2) Steady-State Choice Distribution: Theorem 3 motivates us to compute the asymptotic distribution of y. In this section, we compute the expected steady-state distribution of y by using the expected state transition matrix $\mathbf{P}$. Since the Markov process in Section V-A modeled by the expected state transition matrix $\mathbf{P}$ is irreducible and aperiodic, it has a unique limiting distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_N)$ describing the fraction of time the process will spend in each of the enumerated states $y = i/N$, $i = 0, 1, 2 \ldots N$, in the long run (as $t \to \infty$) [29]. This steady-state distribution is the solution to (14) and (15).

Proposition 5: For the expected transition probabilities given by (23)–(25), the unique expected steady-state distribution is

$$\pi_i = \alpha_i \frac{\prod_{j=1}^{i} q\left(\frac{i}{N}, \nu\right)}{\sum_{j=0}^{N} \alpha_j \prod_{k=1}^{j} q\left(\frac{k}{N}, \nu\right)} \tag{29}$$

where $\alpha_i = N!/((N-i)!i!)$ and $q((i/N), \nu) = \bar{p}_A(((i-1)/N), \nu)/(1 - \bar{p}_A((i/N), \nu))$.

3) A Different Proof for Convergence: We have shown that the distribution of $y(t)$ converges in probability and provided the explicit form for the expectation of the steady-state distribution. The convergence results can be further strengthened by utilizing the fact that $P(t)$ arising in the tested TAFC tasks belongs to a compact set. More specifically, we show here that the elements on the diagonal, superdiagonal, and subdiagonal of $P(t)$ are lower bounded by a positive constant. In the TAFC tasks, we usually have $\mu \in [0, 15]$ and $\nu \in [0, 1]$. Then, from (26)

$$\frac{1 - (1 - p_A(\infty, 0))^{\left\lfloor \frac{M}{2} \right\rfloor - 1}}{1 + e^{\mu(\Delta r + \nu)}} \leq \bar{p}_A(y, \nu) \leq \frac{1 - (1 - p_A(\infty, 0))^{\left\lfloor \frac{M}{2} \right\rfloor - 1}}{1 + e^{\mu(\Delta r - \nu)}}.$$

Let $\kappa_1 = (1 - (1 - p_A(\infty, 0))^{\lfloor M/2 \rfloor - 1})/(1 + e^{\mu(\Delta r + \nu)})$. Since $p_A(\infty, 0) > 0$, we have

$$\bar{p}_A(y, \nu) \geq \kappa_1 > 0. \tag{30}$$

Let $\kappa_2 = (1 - (1 - p_A(\infty, 0))^{\lfloor M/2 \rfloor - 1})/(1 + e^{\mu(\Delta r - \nu)})$, then

$$\bar{p}_A(y, \nu) \leq \kappa_2 < 1. \tag{31}$$

Then, for $y(t) > 0$, from (23)

$$\Pr\left\{y(t+1) = y(t) - \frac{1}{N}\right\} \geq (1 - \kappa_2)\frac{1}{N}. \tag{32}$$

In other words, we have found a positive lower bound for the elements on the subdiagonal of $P(t)$. Similarly, for $y(t) < 1$, from (25)

$$\Pr\left\{y(t+1) = y(t) + \frac{1}{N}\right\} \geq \kappa_1 \frac{1}{N}. \tag{33}$$

So we have constructed a positive lower bound for the elements on the superdiagonal of $P(t)$. When $0 < y(t) < 1$, from (24), we know that $\Pr\{y(t+1) = y(t)\}$ is a convex combination of the values of $1 - y(t)$ and $y(t)$, so

$$\Pr\{y(t+1) = y(t)\} \geq \frac{1}{N}.$$

It is easy to check that when $y(t) = 0$

$$\Pr\{y(t+1) = y(t)\} = 1 - \bar{p}_A(0) \geq 1 - \kappa_2$$

and when $y(t) = 1$

$$\Pr\{y(t+1) = y(t)\} = \bar{p}_A(1) \geq \kappa_1.$$

So

$$\Pr\{y(t+1)=y(t)\}\geq \min\left\{\frac{1}{N},1-\kappa_2,\kappa_1\right\}. \tag{34}$$

Hence, we have also constructed a positive lower bound for the elements on the diagonal of $P(t)$. In view of (32), (33), and (34), we have proved the following result.

Proposition 6: All the elements on the diagonal, subdiagonal, and superdiagonal of $P(t)$ are lower bounded by a positive constant

$$\kappa=\min\left\{(1-\kappa_2)\frac{1}{N},\kappa_1\frac{1}{N},\frac{1}{N},1-\kappa_2,\kappa_1\right\}. \tag{35}$$

Proposition 6 can be used to prove a stronger convergence result for $y(t)$ than that stated in Theorem 3.

Theorem 4: For the TAFC tasks under consideration, there always exists a probability distribution π such that

$$\lim_{t\to\infty} R(t)=\mathbf{1}\pi.$$

The approach taken in this section that uses tools from matrix analysis has one additional advantage: it allows us to conveniently examine the convergence rate by looking at the relevant elements of $P(t)$.

4) Convergence Rate: We first review some results on the estimation of the convergence rate of inhomogeneous Markov chains in the literature. For a stochastic matrix S, its *scrambling constant* [32], [33] is defined to be

$$\varrho(S)=\max_{i,j}\left(1-\sum_{k=1}^{n}\min\{s_{ik},s_{jk}\}\right). \tag{36}$$

A stochastic matrix is then called a *scrambling matrix* if its scrambling constant is strictly less than one. It is well known [32], [33] that the product of any infinite sequence of scrambling matrices $S_1,S_2,\ldots$ from a compact set $\mathcal{S}$ converges exponentially fast to a rank-one matrix at a rate no slower than

$$\max_{S\in\mathcal{S}}\varrho(S).$$

Now we use the scrambling constant to estimate the convergence rate of $y(t)$.

It is easy to check that any stochastic matrix with a positive column is a scrambling matrix. For any sequence of $\lfloor N/2\rfloor$ state transition matrices $P(t),P(t+1),\ldots,P(t+\lfloor N/2\rfloor-1)$, let $R(t)$ denote the matrix product $P(t)P(t+1)\cdots P(t+\lfloor N/2\rfloor-1)$. Consider $j=\lfloor N/2\rfloor+1$. Then, for any $i\in\{1,2,\ldots,N+1\}$, from the tridiagonal structures of the state transition matrices we know that $P(t)_{i,i+1},P(t+1)_{i+1,i+2},\ldots,P(t+j-i-1)_{i,j}$, $P(t+j-i)_{j,j},\ldots,P(t+\lfloor N/2\rfloor-1)_{j,j}$ are all lower bounded by the positive constant κ as defined in (35). Then, the ijth element of $R(t)$ is positive. Since such a conclusion holds for all i, the jth column of $R(t)$ must be positive. Hence, the product of any $\lfloor N/2\rfloor$ state transition matrices $P(t)$ is a scrambling matrix.

The above argument implies the following lower bound for $\varrho(R(t))$:

$$\varrho(R(t))\geq 1-\min\{R_{1,j},R_{N+1,j}\}.$$

It can be further checked that when $u=1$, $R_{1,j}$ increases and $R_{N+1,j}$ decreases and when $u=-1$, $R_{1,j}$ decreases and $R_{N+1,j}$ increases. This suggests that the effect of social feedback on the convergence rate of $y(t)$ is a delicate issue and the design of social feedback for the purpose of accelerating convergence is an interesting and worthwhile topic for future research.

C. Performance With Choice Feedback in the CG Task

A key measure of performance in the TAFC task with CG reward structure is variance about the optimal choice sequence. We consider symmetric CG structures of the form (19) in which $c_A=c_B$, $\bar{y}_A<\bar{y}_B$, and $|\bar{y}_A|=|\bar{y}_B|$ [as in Fig. 1(c)], so that $y=0.5$ is the converging matching point and the optimal choice sequence. Accordingly, better performance corresponds to minimizing the variance about $y=0.5$.

Let Σ denote the variance, or second moment, of the expected steady-state distribution about $y=0.5$. Then, using π given by (29), Σ can be written as a function of the feedback gain ν as

$$\Sigma(\nu)=\frac{\sum_{i=0}^{N}\alpha_i\left(\frac{i}{N}-\frac{1}{2}\right)^2Q_i(\nu)}{\sum_{j=0}^{N}\alpha_jQ_j(\nu)} \tag{37}$$

where $Q_i(\nu):=\prod_{j=1}^{i}q((j/N),\nu)$.

1) Effect of Choice Feedback on Variance About Optimum: We prove here that variance is minimized at $\nu = 0$ in the case $M = 4$. This implies that with choice feedback (corresponding to $\nu \neq 0$), the focal individual tends to do more exploring away from the optimal solution and performance deteriorates.

Theorem 5: Consider the CG reward structure of the form (19) where the matching point and optimal choice sequence coincide at $y = 0.5$. Consider a focal decision maker who receives choice feedback from $M = 4$ others who receive no feedback. Suppose that Assumptions 1 and 2a hold. Then, the variance $\Sigma(\nu)$ about $y = 0.5$ of the expected steady-state choice distribution of the focal individual is minimal for $\nu = 0$.

Evidence of increased exploration and deterioration of performance in the CG task with choice feedback has also been observed in the experimental data of [8], where feedback was undirected. We compute the variance in choice history about $y = 0.5$ as a function of time (choice number from 1 to 150) by averaging over all subjects in the experiments. In Fig. 8, we plot this variance from the experimental data of [8] in the case of undirected choice feedback (dashed curve) and compare to the case of no feedback (solid curve); as predicted by Theorem 5, the experiments show that choice feedback increases variance about $y = 0.5$ for the average subject.

2) Performance Sensitivity to Parameters: We next examine the sensitivity of performance in the CG task with choice feedback to parameters ν, μ, and M. In Fig. 9, the (expected) steady-state distribution of y is plotted with and without feedback in the cases $M = 4$ and $M = 2$. We use $\mu = 2.6$, which is the fitted value for an individual in the CG task with social feedback [8]. In Fig. 10, the normalized standard deviation $100\sqrt{\Sigma(\mu, \nu, M)}$ is plotted as a function of ν for three different values of μ and M.

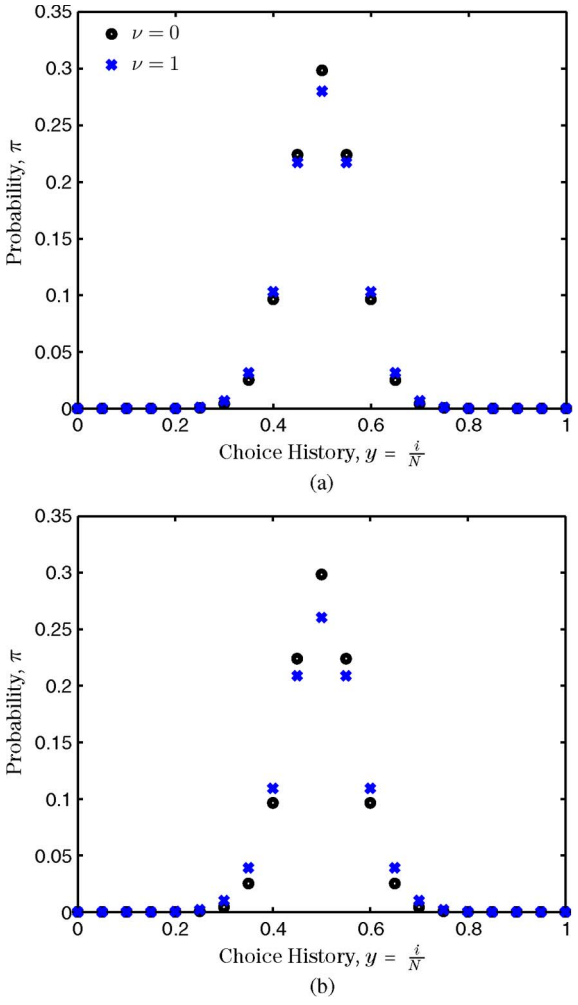

Fig. 9. *Expected steady-state distribution π of y from (29) for the CG task with choice feedback. (a) $M = 4$. (b) $M = 2$. In each plot $\mu = 2.6$, the black "o" symbols correspond to $\nu = 0$, and the blue "x" symbols to $\nu = 1$.*

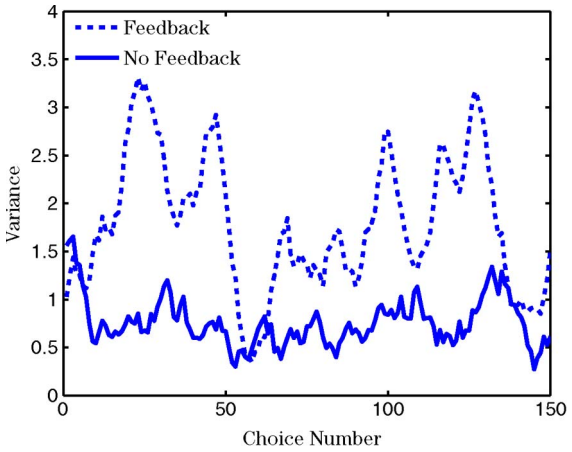

Fig. 8. *Variance in choice history about $y = 0.5$ for experiments in [8]. The case with undirected choice feedback is the dashed curve and with no feedback is a solid curve; both are plotted as a function of time (choice number). $N = 20$.*

In both Figs. 9 and 10, it can be seen that variance increases as a function of ν for each value of μ and M plotted; this is as predicted for the case $M = 4$ by Theorem 5. We also see that variance is higher for smaller M. This implies that choice feedback of this kind has a greater effect on performance in smaller groups.

The results also show that dependence of the variance on μ is significant. In Fig. 10, it can be seen that increasing μ (equivalently, decreasing the exploratory tendency) magnifies sensitivity to the feedback gain ν. In Section IV-D, we showed that increasing μ in the CG task decreases variance for a single individual without social feedback. The exploratory parameter μ and the feedback gain ν have a coupling effect in the CG task with choice

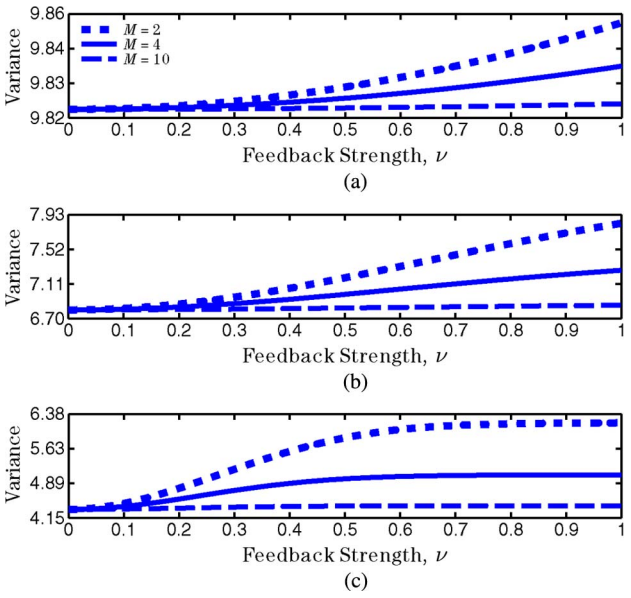

Fig. 10. *Standard deviation of expected steady-state distribution of y from the mean $y = 0.5$ for the CG task as a function of feedback parameter ν [given by $100\sqrt{(\mu, \nu, M)}$]. (a) $\mu = 0.5$. (b) $\mu = 2.6$. (c) $\mu = 10$. In each plot, the short-dashed curve corresponds to $M = 2$, the solid curve to $M = 4$, and the long-dashed curve to $M = 10$.*

feedback that causes a more substantial decrease in performance as ν increases for larger values of μ.

D. Performance in the RO Task With Choice Feedback From Designed Decision Makers

In this section, we use our analytic results to study performance in the RO task with choice feedback. We examine how the ability of a focal decision maker to find the optimal solution in the RO task is influenced by the design of the rest of the team members who provide choice feedback. The space of design alternatives is large; we focus on a parameterized family of designs for systematic evaluation. We let the rest of the team be a heterogeneous group of $M = 4$ decision makers (the same M as in [8]), and we prescribe the (constant) probability $p_{A,m}$ that decision maker m chooses A, for $m = 1, 2, 3, 4$.

To measure performance, we consider the likelihood that the focal decision maker finds the optimal solution at $y = 1$. This we can formalize by defining the probability that, at steady state, the focal decision maker's choice history y will be greater than a critical value $y_c \in [0, 1]$

$$\Pr\left\{y > y_c = \frac{i_c}{N}\right\} = \sum_{i=i_c}^{N+1} \pi_i(\mu, \nu, \Delta r, p_{A,1}, \ldots, p_{A,4}). \quad (38)$$

In Fig. 11(a), we illustrate three choices of y_c as vertical gray lines: $y_c = 0.45, 0.70$, and 0.85 are light, medium, and dark lines, respectively. The probability in (38) is the fraction of the steady-state distribution of y to the right of the vertical line, i.e., the fraction of the distribution in the gray shaded area (light, medium, and dark, respectively).

Because this performance measure is an analytic function of model and design parameters, we can evaluate performance systematically. As design choices, we let $p_{A,1} = 0.05$, implying that decision maker 1 spends a lot of time near the local optimum at $y = 0$, $p_{A,2} = 0.95$, implying that decision maker 2 spends a lot of time near the global optimum at $y = 1$, and $p_{A,3} = 0.5$, implying that decision maker 3 chooses randomly between A and B.

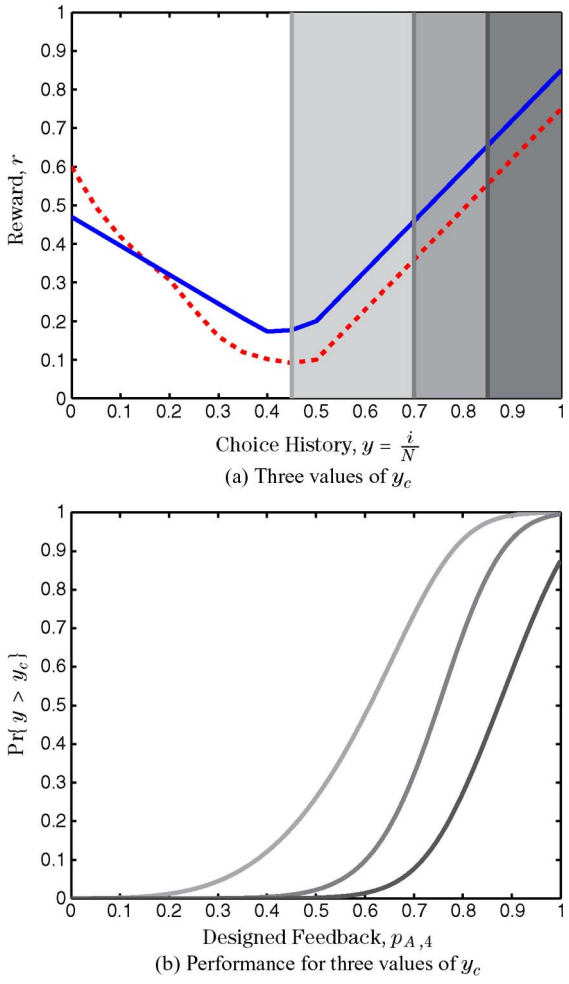

Fig. 11. *Probability given by (38) that at steady state $y > y_c$ in the RO task for a focal decision maker receiving choice feedback from four decision makers with $p_{A,1} = 0.05$, $p_{A,2} = 0.95$, and $p_{A,3} = 0.5$. (a) The values $y_c = 0.45, 0.7$, and 0.85 are shown as (light, medium, and dark) vertical lines on the plot of RO reward structure. For each value of y_c, the probability $y > y_c$ describes the fraction of time that a decision maker will spend with y inside the shaded region to the right of the corresponding vertical line. (b) Probability $y > y_c$ as a function of $p_{A,4}$ for $y_c = 0.45$ (light), $y_c = 0.70$ (medium), and $y_c = 0.85$ (dark).*

The average of these three probabilities is 0.5, which again corresponds to a random decision maker. We look at the sensitivity of performance of the focal decision maker to the design of decision maker 4, parametrized by $p_{A,4}$.

Fig. 11(b) plots the performance from (38) as a function of $p_{A,4}$ for $y_c = 0.45, 0.7$, and 0.85 (light, medium, and dark curves, respectively). The $y_c = 0.45$ curve measures how likely the focal decision maker is to choose A frequently enough to make it past the minimum reward choice sequence. The $y_c = 0.7$ curve measures how likely the focal decision maker is to move well beyond the minimum reward choice sequence. The $y_c = 0.85$ curve measures how likely the focal decision maker is to spend time near the global optimum. The three curves have similar sigmoidal shape with relatively steep slope: performance increases with increasing likelihood $p_{A,4}$ that decision maker 4 chooses A, and this increase is steep after some critical value of $p_{A,4}$. Since higher y_c defines higher performance, higher values of $p_{A,4}$ are required to maintain good performance; thus, the curves move to the right with increasing y_c.

Fig. 12 shows the corresponding expected steady-state distribution for the focal individual with choice feedback from the same four decision makers in cases $p_{A,4} = 0.05$ [Fig. 12(a)] and $p_{A,4} = 0.95$ [Fig. 12(b)]. It can be observed that the focal decision maker is very sensitive to decision maker 4 in these cases: in Fig. 12(a), when decision maker 4 spends most time near the local optimum, so does the focal decision maker, and in Fig. 12(b), when decision maker 4 spends most time near the global optimum, so does the focal decision maker.

Figs. 11 and 12 illustrate how parameters can be chosen in the design of a decision-making team to improve performance of a focal decision maker receiving choice feedback. Experiments with human subjects receiving choice feedback from designed decision makers are underway to test the predictions of Fig. 12.

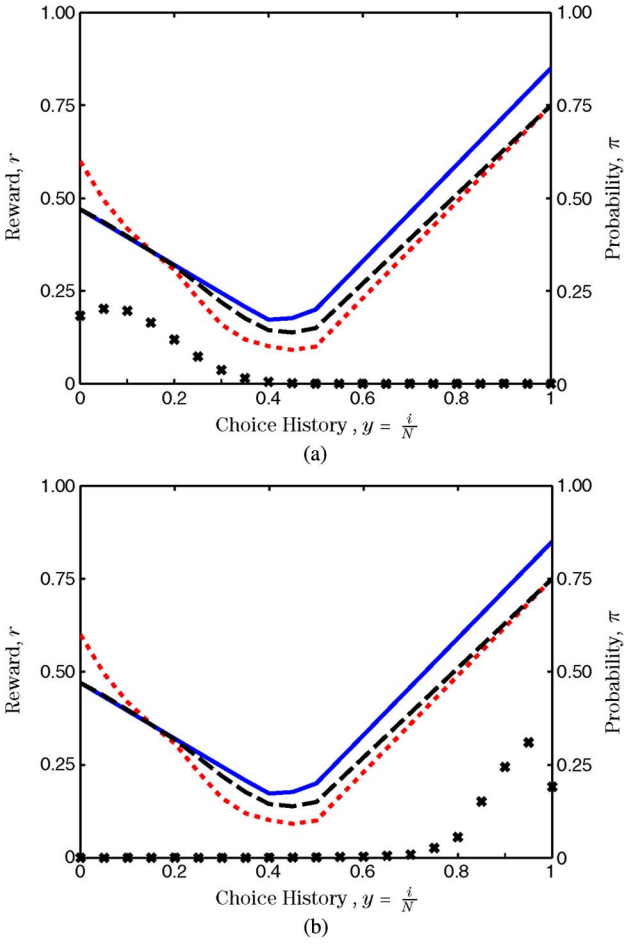

Fig. 12. *Expected steady-state distribution π of y from (29) for a focal decision maker in the RO task, receiving choice feedback from four designed decision makers with $p_{A,1} = 0.05$, $p_{A,2} = 0.95$, and $p_{A,3} = 0.5$. The distribution is plotted with "x" symbols on the RO reward structure. (a) $p_{A,4} = 0.05$. (b) $p_{A,4} = 0.95$.*

E. Undirected Feedback

In this section, we study decision making in the case of undirected choice feedback. We consider again a group of $(M+1)$ model decision makers simultaneously making choices in the TAFC task. However, in the undirected case, each decision maker receives choice feedback from each of the other M decision makers, i.e., the graph that describes the communication topology is complete. The probability that any of the decision makers chooses A is given by (9) where feedback depends on the choices of others. We make Assumptions 1 and 2a (or 2b) so that the state of decision maker k is $y_k(t)$, $k = 1, \ldots, M+1$. Because the decision makers are all interconnected, we must retain the state of each decision maker, so the state of the system becomes $(y_1(t), \ldots, y_{M+1}(t))$.

To study the dynamics, we first identify the task and decision-making model as a Markov process. As in Section V-A, the Markov process is inhomogeneous, and we can compute the expectation of the state transition probabilities by conditioning on $u_k(t)$, $k = 1, \ldots, M+1$. We then use these probabilities to build the expected state transition matrix $\mathbf{P}$, which in this case will be a matrix of dimension $(N+1)^{M+1} \times (N+1)^{M+1}$.

Proposition 7: Suppose Assumptions 1 and 2a hold. Then, $(M+1)$ model decision makers each receiving choice feedback from the M others (9)–(10) for the TAFC task (1)–(3) form a Markov process with state $(y_1(t), \ldots, y_{M+1}(t))$ and expected state transition probabilities given by

$$\Pr\left\{y_i(t+1) = y_i(t) + \frac{d_i}{N}, i=1,\ldots,M+1\right\} = \prod_{i=1}^{M+1} \hat{p}_{i,d_i} \tag{39}$$

where

$$\hat{p}_{m,d}(t) = \begin{cases} \big(1 - p_{A,m}(t)\big) y_m(t) & \text{if } d = -1 \\ p_{A,m}(t) y_m(t) & \\ \quad + \big(1 - y_m(t)\big)\big(1 - p_{A,m}(t)\big), & \text{if } d = 0 \\ p_{A,m}(t)\big(1 - y_m(t)\big), & \text{if } d = 1 \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

The probability $p_{A,m}(t)$ that decision maker m chooses A is given by

$$p_{A,m}(t) = \frac{\Pr\{u_m(t) = 1\}}{1 + e^{\mu_m(\Delta r(y_m) - \nu_m)}} + \frac{\Pr\{u_m(t) = -1\}}{1 + e^{\mu_m(\Delta r(y_m) + \nu_m)}} + \frac{\Pr\{u_m(t) = 0\}}{1 + e^{\mu_m(\Delta r(y_m))}}. \tag{41}$$

$\Pr\{u_m(t) = 1\}$ (respectively, $\Pr\{u_m(t) = -1\}$) is the probability that, among the M decision makers excluding decision maker m, at least $\lceil (M+1)/2 \rceil$ chose A (respectively, B) at time t. $\Pr\{u_m(t) = 0\} = 1 - \Pr\{u_m(t) = 1\} - \Pr\{u_m(t) = -1\}$ is the probability that an equal number of A's and B's were chosen. In case Assumption 2b holds instead of Assumption 2a, then the results hold with $\Delta r(y(t))$ replaced with $f(y(t))$.

Because of the high dimensionality of the matrix $\mathbf{P}$, we compute the expected distributions numerically. This is done by raising $\mathbf{P}$ to a high power so that the elements along each column are equal. Any row in the resulting matrix then has the steady-state distribution as its elements. All rows being equal implies that the probability of transitioning to any of the possible states in the long run is independent of the initial condition.

In Fig. 13, we plot (with "x" symbols) the numerically computed expected steady-state distribution of y for one of the decision makers where there is undirected choice feedback and $M = 2$. We compare this to the plot (with "o" symbols) of the expected steady-state distribution of y from (29) for the focal decision maker in the case of directed choice feedback and $M = 2$. The case of the CG task with $\mu = 2.6$ is shown in Fig. 13(a). We see that there is little difference in the distributions, suggesting that for our model the CG task results do not depend significantly on whether the feedback is undirected or directed. This is consistent with our comparison between the model predictions in the directed case and the experimental data in the undirected case for the CG task as described in Section V-C1.

The case of the DG task with $\mu = 2.9$ is shown in Fig. 13(b). The plot shows that for the DG task the undirected case can differ substantially from the directed case. The focal decision maker makes steady-state choices in the undirected case that are further from the optimal solution as compared to the steady-state choices in the directed case. This suggests that undirected feedback reinforces the tendency for decision makers to move toward relatively high and low values of y, leading to reduced performance as compared to the directed feedback case. The result illustrates the influence that the interconnection topology, in this case directed versus undirected interconnections, can have on the performance of individuals in a group.

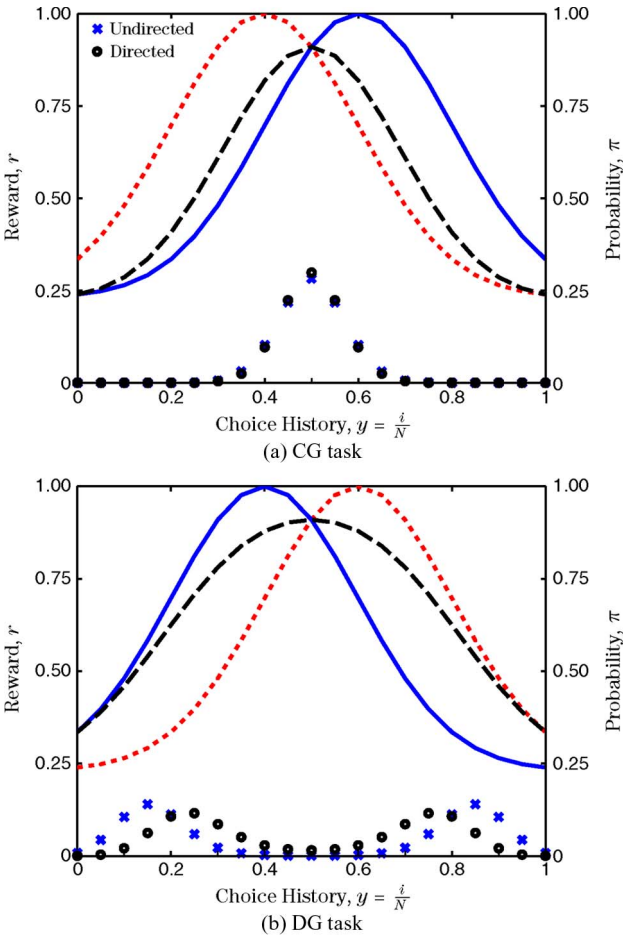

Fig. 13. *Comparison of expected steady-state distribution of y with undirected versus directed choice feedback for $M = 2$. In both plots, black "o" symbols correspond to directed feedback [where π is computed from (29)] and blue "x" symbols correspond to undirected feedback (where π is computed numerically). (a) CG task with $\mu = 2.6$. (b) DG task with $\mu = 2.9$.*

VI. CONCLUSION AND FUTURE WORK

We have studied decision making in the TAFC task using the empirically verified soft-max choice model and the new extension of this model to decision making in the social context [8], [19]. We have derived analytic expressions that predict steady-state decision sequences and performance of individuals in a group who make

choices at the same time in the same TAFC task in response to feedback on their own performance and on the choices made by others in the group. From these derived expressions, performance can be systematically evaluated as a function of parameters; alternatively, parameters can be systematically selected to meet desired performance criteria. This provides an important step towards a principled approach to design of decision-making groups.

The derived expressions for performance depend explicitly on parameters associated with the task, the decision makers, and the feedback interconnections. Parameters that define the task include the reward curves, r_A and r_B, and the number of past choices N upon which the reward depends. Each decision maker is defined by a certainty parameter μ, which models the tendency to explore, and a feedback gain ν, which models how much attention is paid to the choice feedback received. Other key parameters include M, the number of individuals providing choice feedback, and the topology of the feedback network, e.g., undirected versus directed choice feedback.

If the decision-making group is composed of robots as well as humans, and the robots make choices according to the soft-max choice model, then the parameters that define the composition of the team, the defining parameters for the robots, e.g., μ and ν, and the network topology can be designed to improve performance. The results on convergence in Section V-B suggest further possibilities for design of social feedback to maximize convergence rate in decision making. We have explored sensitivity of performance in four prototypical TAFC tasks to many of these parameters. We have also illustrated the design of a team in which one human decision maker receives feedback from four other programmed decision makers in the challenging RO task; experimental testing of this design is underway.

Future work includes using the model to further explore a range of stationary and time-varying reward structures, heterogeneous groups of decision makers, the role of alternative interconnection topologies, and open questions in convergence and convergence rates. It is also of interest to extend our methods to other types of social feedback as well as framing effects, such as the incentive structure provided to participants and the identification of group members as human or robotic.

Future work also includes experiments with a multirobot testbed and a multihuman interface to explore the applicability of our results to real-world scenarios. New experiments are planned that leverage the generalizability of the TAFC tasks and extend our framework to human–robot team decision making in tasks that require balancing exploration and exploitation to search noisy, unknown, spatially distributed resource fields.

First, in the case that a mixed group of human and robot decision makers is assigned to a decision-making problem that maps to a TAFC task, then the results of the present paper can be applied directly to the design of high-performing decision-making teams. For example, if the human members of the team can be selected, design parameters include the number of humans with strong exploratory tendencies (e.g., risk takers) and the number of humans with more conservative tendencies. If each robot is programmed to make decisions according to the same model, then design parameters include the number of robots and the value of μ and ν for each robot. The network topology, i.e., who receives feedback from whom, can also be designed for good performance.

To make this more concrete, consider the following example of a decision-making problem that maps onto the class of TAFC tasks. The setting is the Gulf of Mexico after the BP oil spill where two autonomous vehicles move around in a fixed region, each making a different, regular pattern just below the water surface. Each vehicle measures concentration of oil, recording its position with every sample. A land-based operator seeks to acquire data on locations of high oil concentration in real time in order to more quickly aid oil cleanup activities. The operator can query the vehicles for data readings at regular time intervals; however, because of bandwidth limitations, only one of the two vehicles can be queried at a time. Thus, the operator must choose between vehicle A and vehicle B at every time interval; in response, the operator will get a score on the merit of the latest query. The score reflects the value of the new data to an automated assignment of oil cleanup resources, and it depends on the current choice (high concentration values are helpful) as well as the recent history of choices (redundant data are wasteful).

The decision-making problem is complex due to the combination of the advection of the oil in the water and the dynamics of the vehicles. The complexity implies that the structure of the score may not look as clean as one of the four TAFC reward structures studied in this paper. However, there are likely to be choice sequences that provide high performance, and likewise other attributes of the four prototypical structures, such as a matching point that is not coincident with optimal performance. Further, in the case that this scenario is repeated in several fixed regions, over which there is likely to be correlation in the dynamics of the oil, feedback among operators can prove useful. If we allow some of the operators to be humans and some to be robots, then we can apply the approach described above to design the decision-making team.

Beyond the problems that map to TAFC tasks (see also [24]), significant potential for application lies in the generalizability of the TAFC tasks. The TAFC tasks span a range of explore-versus-exploit challenges that are representative of fundamental challenges in any decision-making process in an uncertain environment. Accordingly, the results in this paper provide a stepping stone towards a systematic, principled approach for designing decision-making teams for more general decision-making processes that require strategies for explore versus exploit.

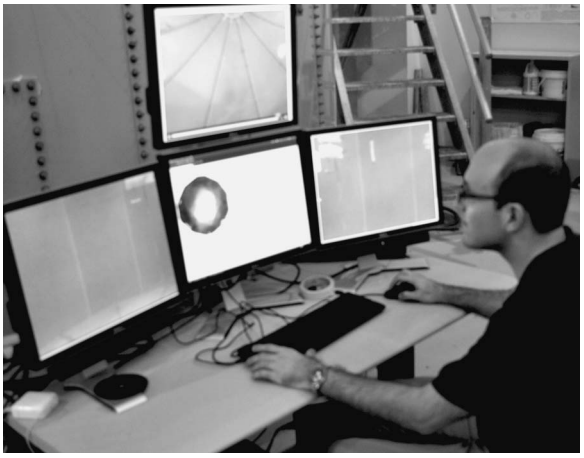

Fig. 14. *Human interface to robotic vehicles in tank testbed. The human can make choices, which are communicated to the vehicle control system, and can observe vehicle performance, including live video feeds.*

Indeed, in ongoing work, we are building on the present paper to examine mixed human–robot decision making in a task where decision makers search for peaks in an unknown, noisy, spatially distributed resource field. At regular time intervals, each decision maker chooses a new location to search. Choices in a new area of the field suggest exploration, while choices in the immediate area suggest exploitation. Social feedback can come in the form of choices (where others are searching) or of rewards (how much resource others are finding). As an example application, consider again the oil field in the Gulf of Mexico. The search problem is one in which each (human or robot) decision maker picks waypoints for a vehicle or coordinated group of vehicles with the goal of finding the location with the highest concentration of oil.

To further investigate the applicability of our work to concrete tasks, we have planned experiments in our 3-D multirobot testbed with human–robot decision-making teams performing tasks much like the two oil spill sampling problems. The testbed can be manipulated to resemble real-world settings and can be quickly reconfigured for an array of conditions. The facility, part of Princeton University's Dynamical Control Systems Laboratory, houses a 20 000-gal water tank and a fleet of small, neutral-buoyancy, submersible vehicles. Access to real-time tracking measurements of the robotic vehicles from the streaming video of a system of overhead cameras adds versatility. For example, the topology of communication among the robots can be designed and any virtual, spatially distributed resource field can be imposed. The human interface, shown in Fig. 14, allows for applications such as human assignment of waypoints in search problems. Importantly, the interface uses the internet for communications so the humans can be remotely located, e.g., in the cognitive psychology laboratory for rigorous human subject experiments. ■

APPENDIX

A. Proof of Proposition 1

Since for a given choice $x_1(t+1)$ at time $t+1$, $y(t+1)$ can only change from its current value of $y(t)$ to $y(t)+(1/N)$, $y(t)-(1/N)$ or stay at $y(t)$, we need only compute the probability of each of these three events for all $y(t) \in \mathcal{Y}$. Each of these events depends on the current value of $y(t)$ as well as $x_1(t+1)$ and $x_N(t)$ since $y(t+1)$ will only differ from $y(t)$ if $x_1(t+1)$ also differs from $x_N(t)$.

The event that $y(t+1) = y(t) - (1/N)$ requires $x_1(t+1) = B$ and $x_N(t) = A$. We treat the latter as independent events since, in these tasks, the decision makers are not told that their reward depends on choice history. Using (5) with Assumption 1 yields

$$\begin{aligned}\Pr\left\{y(t+1) = y(t) - \frac{1}{N}\right\} &= \Pr\{x_1(t+1) = B\} \\ &\quad \times \Pr\{x_N(t) = A\} \\ &= \frac{e^{\mu(w_B(t) - w_A(t))} y(t)}{1 + e^{\mu(w_B(t) - w_A(t))}}.\end{aligned}$$

Substituting in the identity of Assumption 2a, we get (11).

Similarly, the probability that $y(t+1)$ takes the value $y(t) + (1/N)$ is given by

$$\begin{aligned}\Pr\left\{y(t+1) = y(t) + \frac{1}{N}\right\} &= \Pr\{x_1(t+1) = A\} \\ &\quad \times \Pr\{x_N(t) = B\} \\ &= \frac{1 - y(t)}{1 + e^{\mu(w_B(t) - w_A(t))}}.\end{aligned}$$

Substituting in the identity of Assumption 2a, we get (13).

The event that $y(t+1) = y(t)$ requires either $x_1(t+1) = A$ and $x_N(t) = A$ or $x_1(t+1) = B$ and $x_N(t) = B$. The probability of the union of these events is

$$\begin{aligned}&\Pr\{y(t+1) = y(t)\} \\ &\quad = \Pr\{x_1(t+1) = A\} \Pr\{x_N(t) = A\} \\ &\qquad + \Pr\{x_1(t+1) = B\} \Pr\{x_N(t) = B\} \\ &\quad = \frac{y(t) + (1 - y(t)) e^{\mu(w_B(t) - w_A(t))}}{1 + e^{\mu(w_B(t) - w_A(t))}}.\end{aligned}$$

Substituting in the identity of Assumption 2a, we get (12). Since all of the probabilities depend on $y(t)$ only, the state at time t, the process is Markov. The case when Assumption 2b holds follows similarly. ■

B. Proof of Proposition 2

Solving (14) alone yields a row vector v with elements

$$v_i = \frac{N!}{(N-i)!i!}\left(1+e^{\mu\Delta r\left(\frac{i}{N}\right)}\right)e^{-\mu\sum_{j=1}^{i}\Delta r\left(\frac{j}{N}\right)}.$$

To solve (15), we normalize the vector v to get

$$\pi = \frac{v}{\sum_{i=0}^{N} v_i}.$$

The elements of π are then given by (16). ■

C. Proof of Theorem 1

To prove Theorem 1, we examine $\rho(i) = \pi_i/\pi_{i^*}$, the ratio of time spent at $y = i/N$, $i \neq i^*$, to time spent at $y^* = i^*/N$. From (16), we compute

$$\rho(i) = \frac{(N-i^*)!i^*!\left(1+e^{\mu\Delta r\left(\frac{i}{N}\right)}\right)e^{-\mu\sum_{j=1}^{i}\Delta r\left(\frac{j}{N}\right)}}{2(N-i^*)!i^*!e^{-\mu\sum_{j=1}^{i^*}\Delta r\left(\frac{j}{N}\right)}}.$$

We show that $\rho(i) < 1$ for all $i \notin \{i^*-1, i^*, i^*+1\}$ by proving each of two cases. In the first case, we show that $\rho(i) < 1$ for all $i > i^*+1$. In the second case, we show that $\rho(i) < 1$ for $i < i^*-1$.

Case 1: Let $\epsilon = i - i^*$ with $\epsilon > 0$. Then, we have

$$\rho(i) = \frac{(N-i^*)!i^*!\left(1+e^{\mu\Delta r\left(\frac{i^*+\epsilon}{N}\right)}\right)e^{-\mu\left(\Delta r\left(\frac{i^*+1}{N}\right)+\cdots+\Delta r\left(\frac{i^*+\epsilon}{N}\right)\right)}}{2(N-i^*-\epsilon)!(i^*+\epsilon)!}. \tag{42}$$

Replacing $(N-i)!i!$ in the denominator of (42) with its minimal possible value for $i \in \{0,1,\ldots,N\}$ yields

$$\rho(i) \leq \gamma\left(1+e^{-\mu\Delta r\left(\frac{i^*+\epsilon}{N}\right)}\right)e^{-\mu\left(\Delta r\left(\frac{i^*+1}{N}\right)+\cdots+\Delta r\left(\frac{i^*+\epsilon-1}{N}\right)\right)} \tag{43}$$

where $\gamma = ((N-i^*)!i^*!)/(2\lfloor N/2\rfloor!\lceil N/2\rceil!)$.

Now assume $\epsilon \geq 2$. Since $\Delta r((i^*+\epsilon)/N) > 0$ for all $\epsilon \geq 1$, $\rho(i)$ decreases with increasing ϵ so from (43) we can write

$$\rho(i) < \gamma\left(1+e^{-\mu\Delta r\left(\frac{i^*+2}{N}\right)}\right). \tag{44}$$

If (17) is satisfied, then (44) becomes $\rho(i) < 1$.

Case 2: Let $\epsilon = i - i^*$ with $\epsilon < 0$. Following the same steps as in Case 1, and making use of the fact that $\Delta r((i^*-\epsilon)/N) < 0$ for all $\epsilon > 0$, we can write

$$\rho(i) \leq \gamma\left(1+e^{-\mu\left|\Delta r\left(\frac{i^*-\epsilon}{N}\right)\right|}\right)e^{-\mu\left(\left|\Delta r\left(\frac{i^*-\epsilon+1}{N}\right)\right|+\cdots+\left|\Delta r\left(\frac{i^*}{N}\right)\right|\right)}.$$

Now assume $\epsilon \leq -2$. Since $\rho(i)$ decreases with decreasing ϵ for $\epsilon < 0$, we can write

$$\rho(i) < \frac{(N-i^*)!i^*!}{\lfloor\frac{N}{2}\rfloor!\lceil\frac{N}{2}\rceil!}\left(1+e^{-\mu\left|\Delta r\left(\frac{i^*-2}{N}\right)\right|}\right). \tag{45}$$

If (17) is satisfied, then (45) becomes $\rho(i) < 1$. ■

D. Proof of Theorem 2

Again we examine $\rho(i) = \pi_i/\pi_{i^*}$. To prove Theorem 2, we follow the same process used in Theorem 1. We show that $\rho(i) < 1$ for all $i \neq i^*$ by proving each of two cases. In the first case, we show that $\rho(i) < 1$ for all $i > i^*$. In the second case, we show that $\rho(i) < 1$ for $i < i^*$.

Case 1: Let $\epsilon = i - i^*$ with $\epsilon > 0$. Assume $\epsilon \geq 1$. We have shown that $\rho(i)$ decreases with increasing ϵ so using (43), we arrive at

$$\rho(i) < \frac{N-i^*}{2(i^*+1)}\left(1+e^{-\mu\Delta r\left(\frac{i^*+1}{N}\right)}\right). \tag{46}$$

If (18) is satisfied, then (46) becomes $\rho(i) < 1$.

Case 2: Let $\epsilon = i - i^*$ with $\epsilon < 0$. We assume $\epsilon \leq -1$. Since $\rho(i)$ decreases with decreasing ϵ for $\epsilon < 0$, then

$$\rho(i) < \frac{(N-i^*)!i^*!}{\lfloor\frac{N}{2}\rfloor!\lceil\frac{N}{2}\rceil!}\left(1+e^{-\mu\left|\Delta r\left(\frac{i^*-1}{N}\right)\right|}\right). \tag{47}$$

If (18) is satisfied, then (47) becomes $\rho(i) < 1$. ■

E. Proof of Proposition 3

Since for a given choice $x_1(t+1)$ at time $t+1$, $y(t+1)$ can only change from its current value of $y(t)$ to $y(t)+(1/N)$, $y(t)-(1/N)$ or stay at $y(t)$, we need only compute the probability of each of these three events for all $y(t) \in \mathcal{Y}$. Each of these events depends on the current

value of $y(t)$ as well as $x_1(t+1)$ and $x_N(t)$ since $y(t+1)$ will only differ from $y(t)$ if $x_1(t+1)$ also differs from $x_N(t)$.

The event that $y(t+1) = y(t) - (1/N)$ requires $x_1(t+1) = B$ and $x_N(t) = A$. Treating these as independent events and using (9) yields

$$\begin{aligned}\Pr\left\{y(t+1) = y(t) - \frac{1}{N}\right\} &= \Pr\{x_1(t+1) = B\}\Pr\{x_N(t) = A\} \\ &= \frac{e^{\mu(w_B(t)-w_A(t)-\nu u(t))}y(t)}{1+e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}.\end{aligned}$$

We treat the M peer decisions as independent events since in these tasks the decision makers are told that their performance does not depend on the choices or rewards of others. Substituting in the identity of Assumption 2a, we condition on the value of $u(t)$ and get $\Pr\{x_1(t+1) = B\} = 1 - \bar{p}_A(y(t), \nu)$, which with Assumption 1 gives us (23).

Similarly, the probability that $y(t+1) = y(t) + (1/N)$ is

$$\begin{aligned}\Pr\left\{y(t+1) = y(t) + \frac{1}{N}\right\} &= \Pr\{x_1(t+1) = A\}\Pr\{x_N(t) = B\} \\ &= \frac{1-y(t)}{1+e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}.\end{aligned}$$

Conditioning on the value of $u(t)$ and substituting in the identity of Assumption 2a, we get (25).

The event that $y(t+1) = y(t)$ requires either $x_1(t+1) = A$ and $x_N(t) = A$ or $x_1(t+1) = B$ and $x_N(t) = B$. The probability of the union of these events is

$$\begin{aligned}&\Pr\{y(t+1) = y(t)\} \\ &\quad = \Pr\{x_1(t+1) = A\}\Pr\{x_N(t) = A\} \\ &\qquad + \Pr\{x_1(t+1) = B\}\Pr\{x_N(t) = B\} \\ &\quad = \frac{y(t) + (1-y(t))e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}{1+e^{\mu(w_B(t)-w_A(t)-\nu u(t))}}.\end{aligned}$$

Conditioning on the value of $u(t)$ and substituting in the identity of Assumption 2a, we get (24).

Since the probabilities depend only on $y(t)$, the current value of the state at time t, the process is Markov. By conditioning on $u(t)$, the results (23)–(25) provide the expectation of the transition probabilities. The case when Assumption 2b holds follows similarly. ■

F. Proof of Theorem 3

Consider $i, j \in \{1, \ldots, N+1\}$. Without loss of generality, we assume $i \leq j$ and let $l = j - i - 1$. Since the elements on the superdiagonals of $P(t)$ are positive, we know $P(0)_{i,i+1}, P(1)_{i+1,i+2}, \ldots, P(l)_{j-1,j} > 0$. Then, $R(l)_{ij} > P(0)_{i,i+1}P(1)_{i+1,i+2}\cdots P(l)_{j-1,j} > 0$ and thus $\Pr\{R(l)_{ij} > 0\} = 1$. Similarly, since the elements on the diagonals of $P(t)$ are positive, we know that $R(l)_{jj} > P(0)_{jj}P(1)_{jj}\cdots P(l)_{jj} > 0$ and thus $\Pr\{R(l)_{jj} > 0\} = 1$. Applying Proposition 4 by taking $k = j$ and $t = l$, we know that as t goes to infinity, $R(t)$ converges to a rank-one matrix almost surely and thus π always exists. ■

G. Proof of Proposition 5

Solving (14) alone yields a row vector v whose elements are given by

$$v_i = \frac{N!}{(N-i)!i!}\prod_{j=1}^{i}\frac{\bar{p}_A\left(\frac{j-1}{N}, \nu\right)}{1-\bar{p}_A\left(\frac{j}{N}, \nu\right)}.$$

To solve (15), we normalize the vector v to get $\pi = v/\sum_{i=0}^{N} v_i$. Then, π is given by (29). ■

H. Proof of Theorem 4

From Proposition 6, each $P(t)$ is an ergodic stochastic matrix [32]. To be more precise, each $P(t)$ is taken from a compact set of ergodic stochastic matrices with nonzero elements lower bounded by a positive constant κ. Then, from classical results on inhomogeneous Markov chains [32], the product of any infinite sequences of such $P(t)$ will always converge to a rank-one matrix. ■

I. Proof of Theorem 5

We prove Theorem 5 by first proving four lemmas.

Lemma 1: $\nu = 0$ is a critical point of $\Sigma(\nu)$

To prove Lemma 1, we introduce the following.

Lemma 2: $Q_i'(0) := (\partial/\partial\nu)\prod_{j=1}^{i} q((i/N), \nu)\big|_{\nu=0} = 0$.

Proof of Lemma 2: We compute

$$\begin{aligned}\frac{\partial}{\partial\nu}q\left(\frac{i}{N}, \nu\right) &= \frac{\frac{\partial}{\partial\nu}\bar{p}_A\left(\frac{i-1}{N}, \nu\right)}{1-\bar{p}_A\left(\frac{i}{N}, \nu\right)} \\ &\quad + \frac{\frac{\partial}{\partial\nu}\bar{p}_A\left(\frac{i}{N}, \nu\right)\left(1-\bar{p}_A\left(\frac{i-1}{N}, \nu\right)\right)}{\left(1-\bar{p}_A\left(\frac{i}{N}, \nu\right)\right)^2}.\end{aligned}$$

For the CG reward structure $p_A(\infty, 0) = 1/2$. Using this and $M = 4$ in (27), we can compute the conditional probabilities on $u(t)$. Substituting into (26) for $\bar{p}_A$ gives

$$\bar{p}_A\left(\frac{i}{N}, \nu\right) = \frac{3}{4}\frac{1}{(1+e^{\mu\Delta r})} + \frac{1}{8}\left[\frac{1}{1+e^{\mu(\Delta r-\nu)}} + \frac{1}{1+e^{\mu(\Delta r+\nu)}}\right]$$

where $\Delta r = \Delta r(i/N)$. Then

$$\frac{\partial}{\partial\nu}\bar{p}_A\left(\frac{i}{N}, \nu\right) = \frac{\mu e^{\mu\Delta r}}{8}\left[\frac{e^{-\mu\nu}}{\left(1+e^{\mu(\Delta r-\nu)}\right)^2} - \frac{e^{\mu\nu}}{\left(1+e^{\mu(\Delta r+\nu)}\right)^2}\right]. \tag{48}$$

Evaluating (48) at $\nu = 0$, we get $(\partial/\partial\nu)\bar{p}_A((i/N), \nu)|_{\nu=0} = 0$, $\forall i$. Therefore, $(\partial/\partial\nu)q((i/N), \nu)|_{\nu=0} = 0$. Using the definition of $Q_i(\nu)$ from (37), we can write

$$Q_i'(\nu) = \sum_{k=1}^{i}\frac{\partial}{\partial\nu}q\left(\frac{k}{N}, \nu\right)\prod_{j=1, j\neq k}^{i} q\left(\frac{j}{N}, \nu\right). \tag{49}$$

Evaluating (49) at $\nu = 0$ with $(\partial/\partial\nu)q((i/N), \nu)|_{\nu=0} = 0$ gives $Q_i'(0) = 0$. ■

Proof of Lemma 1: The derivative of $\Sigma(\nu)$ is

$$\frac{\partial}{\partial\nu}\Sigma(\nu) = \frac{\sum_{i=1}^{N}\alpha_i\left(\frac{i}{N}-\frac{1}{2}\right)^2 Q_i'(\nu)}{\sum_{k=1}^{N}\alpha_k Q_k(\nu)} - \frac{\sum_{i=1}^{N}\alpha_i\left(\frac{i}{N}-\frac{1}{2}\right)^2 Q_i(\nu)\sum_{k=1}^{N}\alpha_k Q_k'(\nu)}{\left(\sum_{k=1}^{N}\alpha_k Q_k(\nu)\right)^2}.$$

It follows from Lemma 2 that $(\partial/\partial\nu)\Sigma(\mu, \nu)|_{\nu=0} = 0$. ■

It is now left to show that $\nu = 0$ is a minimum of $\Sigma(\nu)$.

Lemma 3: $(\partial^2/\partial\nu^2)\Sigma(\nu)|_{\nu=0} > 0$.

To prove Lemma 3, we introduce the following.

Lemma 4: $Q_i''(\nu) < 0$.

Proof of Lemma 4: Differentiating $Q_i'(\nu)$ with respect to ν, and using the fact that $(\partial/\partial\nu)\bar{p}_A((i/N), \nu)|_{\nu=0} = 0$ gives

$$Q_i''(0) = \frac{\frac{\partial^2}{\partial\nu^2}\bar{p}_A\left(\frac{i-1}{N}, \nu\right)\Big|_{\nu=0}\left(1-\bar{p}_A\left(\frac{i}{N}, 0\right)\right)}{\left(1-\bar{p}_A\left(\frac{i}{N}, 0\right)\right)^2} + \frac{\frac{\partial^2}{\partial\nu^2}\bar{p}_A\left(\frac{i}{N}, \nu\right)\Big|_{\nu=0}\left(\bar{p}_A\left(\frac{i-1}{N}, 0\right)\right)}{\left(1-\bar{p}_A\left(\frac{i}{N}, 0\right)\right)^2}.$$

Since

$$\frac{\partial^2}{\partial\nu^2}\bar{p}_A\left(\frac{i}{N}, \nu\right)\Big|_{\nu=0} = -\frac{\mu^2 e^{\mu\Delta r\left(\frac{i}{N}\right)}\left(1+e^{2\mu\Delta r\left(\frac{i}{N}\right)}\right)}{\left(1+e^{\mu\Delta r\left(\frac{i}{N}\right)}\right)^4} < 0$$

we can conclude that $Q_i''(0) < 0$. ■

Proof of Lemma 3: Invoking Lemma 2, we can write

$$\frac{\partial^2}{\partial\nu^2}\Sigma(\nu)|_{\nu=0} = \frac{\sum_{i=0}^{N}\alpha_i\left(\frac{i}{N}-\frac{1}{2}\right)^2 Q_i''(0)}{\sum_{k=0}^{N}\alpha_k Q_k(0)} - \frac{\sum_{i=0}^{N}\alpha_i\left(\frac{i}{N}-\frac{1}{2}\right)^2 Q_i(0)\sum_{k=0}^{N}\alpha_k Q_k''(0)}{\left(\sum_{k=0}^{N}\alpha_k Q_k(0)\right)^2}.$$

Denote the numerator of $(\partial^2/\partial\nu^2)|_{\nu=0}\Sigma(\nu)$ by Γ. Then

$$\Gamma = \sum_{i=0}^{N}\sum_{k=0}^{N}\gamma_{i,k} \tag{50}$$

where $\gamma_{i,k} = \alpha_i Q_i''(0)\alpha_k Q_k(0)\ [((i/N)-(1/2))^2 - ((k/N)-(1/2))^2]$. Lemma 4 tells us that $\gamma_{i,k} > 0$ for all i, k that satisfy

$$\left(\frac{i}{N}-\frac{1}{2}\right)^2 - \left(\frac{k}{N}-\frac{1}{2}\right)^2 < 0.$$

It is also true that $\gamma_{(N/2),(N/2)} = 0$.

It can be shown for all $i, k \neq (N/2)$ that $\gamma_{i,k} > 0$ and $\gamma_{(N/2),(N/2)} = 0$. It therefore must be true that $\Gamma = \sum_{i=0}^{N} \sum_{k=0}^{N} \gamma_{i,k} > 0$. ■

Proof of Theorem 5: Lemma 1 and Lemma 3 guarantee that $\nu = 0$ is a minimum of $\Sigma(\nu)$. ■

J. Proof of Proposition 7

Since for a given choice by decision maker m at time $t+1$, $y_m(t+1)$ can only change from its current value of $y_m(t)$ to $y_m(t) + (1/N)$, $y_m(t) - (1/N)$, or stay at $y_m(t)$, we need only compute the probability $\hat{p}_{m,d}$ of each of these three events, $d = 1, d = -1, d = 0$, for all $y_m(t) \in \mathcal{Y}$ and each m. Each of these events depends on the current state $(y_1(t), \ldots, y_M(t))$, as well as each decision maker's most recent choice and oldest choice in their history of N choices, since $y_m(t+1)$ will only differ from $y_m(t)$ for decision maker m if the most recent decision also differs from the oldest decision in the history. The probabilities $\hat{p}_{m,d}$ of (40) are derived analogously to (23)–(25) in Proposition 3 with the probability $p_{A,m}$ that decision maker m chooses A of (41) derived analogously to $\bar{p}_A$ of (26). The computation of $p_{A,m}(t)$ requires conditioning on the value of $u_m(t)$.

Treating each decision maker's choice as an independent event, the transition probabilities for the group are given by (39). Since the probabilities depend only on $(y_1(t), \ldots, y_M(t))$, the current value of the state at time t, the process is Markov. The case when Assumption 2b holds follows similarly. ■

REFERENCES

[1] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. E. Davis, "Collective motion, sensor networks, and ocean sampling," *Proc. IEEE*, vol. 95, no. 1, pp. 48–74, Jan. 2007.

[2] R. R. Murphy, "Human-robot interaction in rescue robotics," *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 34, no. 2, pp. 138–153, May 2004.

[3] D. J. Bruemmer, J. L. Marble, D. D. Dudenhoeffer, M. O. Anderson, and M. D. McKay, "Mixed-initiative control for remote characterization of hazardous environments," in *Proc. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Waikoloa, HI, 2003, p. 9.

[4] T. Fong, I. Nourbakshsh, C. Kunz, L. Fluckiger, J. Schreiner, R. Ambrose, R. Burridge, R. Simmons, L. M. Hiatt, A. Schultz, J. G. Trafton, M. Bugajska, and J. Scholtz, "The peer-to-peer human-robot interaction project," in *Proc. AIAA Space*, 2005, pp. 2005–6750.

[5] D. J. Bruemmer, D. A. Few, R. L. Boring, J. L. Marble, M. C. Walton, and C. W. Nielsen, "Shared understanding for collaborative control," *IEEE Trans. Syst. Man Cybern. A, Syst. Humans*, vol. 35, no. 4, pp. 494–504, Jul. 2005.

[6] T. Kaupp, A. Makarenko, and H. Durrant-Whyte, "Human-robot communication for collaborative decision making—A probabilistic approach," *Robot. Autonom. Syst.*, vol. 58, pp. 444–456, 2010.

[7] D. M. Egelman, C. Person, and P. R. Montague, "A computational role for dopamine delivery in human decision-making," *J. Cogn. Neurosci.*, vol. 10, pp. 623–630, 1998.

[8] A. Nedic, D. Tomlin, P. Holmes, D. A. Prentice, and J. D. Cohen, "A decision task in a social context: Human experiments, models, and analyses of behavioral data," *Proc. IEEE*, vol. 100, no. 3, pp. 713–733, Mar. 2012, DOI: 10.1109/JPROC.2011.2166437.

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

[10] R. J. Herrnstein, "Melioration as behavioral dynamism," in *Quantitative Analyses of Behavior, Vol. 2: Matching and Maximizing Account*, M. L. Commons, R. J. Herrnstein, and H. Rachlin, Eds. Cambridge, MA: Ballinger, 1982, pp. 433–458.

[11] R. J. Herrnstein, "Rational choice theory: Necessary but not sufficient," *Amer. Psychol.*, vol. 45, pp. 356–367, 1990.

[12] R. J. Herrnstein, "Experiments on stable suboptimality in individual behavior," *AEA Papers Proc.*, vol. 81, pp. 360–364, 1991.

[13] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks," *Psychol. Rev.*, vol. 113, pp. 700–765, 2006.

[14] R. Ratcliff, "A theory of memory retrieval," *Psychol. Rev.*, vol. 85, pp. 59–108, 1978.

[15] R. Ratcliff, T. Van Zandt, and G. McKoon, "Connectionist and diffusion models of reaction time," *Psychol. Rev.*, vol. 106, no. 2, pp. 261–300, 1999.

[16] P. L. Smith and R. Ratcliff, "Psychology and neurobiology of simple decisions," *Trends Neurosci.*, vol. 27, no. 3, pp. 161–168, 2004.

[17] A. Wald, *Sequential Analysis*. New York: Wiley, 1947.

[18] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Stat.*, vol. 19, pp. 326–339, 1948.

[19] A. Nedic, D. Tomlin, P. Holmes, D. A. Prentice, and J. D. Cohen, "A simple decision task in a social context: Experiments, a model, and preliminary analyses of behavioral data," in *Proc. 47th IEEE Conf. Decision Control*, Cancun, Mexico, 2008, pp. 1115–1120.

[20] A. Stewart, M. Cao, and N. E. Leonard, "Steady-state distributions for human decisions in two-alternative choice tasks," in *Proc. Amer. Control Conf.*, Baltimore, MD, 2010, pp. 2378–2383.

[21] A. Stewart and N. E. Leonard, "The role of social feedback in steady-state performance of human decision making for two-alternative choice tasks," in *Proc. 49th IEEE Conf. Decision Control*, Atlanta, GA, 2010, pp. 3796–3801.

[22] 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 Res.*, vol. 1153, pp. 111–121, 2007.

[23] M. Cao, A. Stewart, and N. E. Leonard, "Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis," in *Proc. 47th IEEE Conf. Decision Control*, Cancun, Mexico, 2008, pp. 1127–1132.

[24] M. Cao, A. Stewart, and N. E. Leonard, "Convergence in human decision-making dynamics," *Syst. Control Lett.*, vol. 59, pp. 87–97, 2010.

[25] L. Vu and K. Morgansen, "Modeling and analysis of dynamic decision making in sequential two-choice tasks," in *Proc. 47th IEEE Conf. Decision Control*, Cancun, Mexico, 2008, pp. 1121–1126.

[26] P. R. Montague, P. Dayan, and T. J. Sejnowski, "A framework for mesencephalic dopamine systems based on predictive Hebbian learning," *J. Neurosci.*, vol. 16, pp. 1936–1947, 1996.

[27] B. K. Oksendal, *Stochastic Differential Equations: An Introduction With Applications*. Berlin, Germany: Springer-Verlag, 2003.

[28] P. Simen and J. D. Cohen, "Explicit melioration by a neural diffusion model," *Brain Res.*, vol. 1299, pp. 95–117, 2009.

[29] H. M. Taylor and S. Karlin, *An Introduction to Stochastic Modeling*, 3rd ed. New York: Academic, 1998.

[30] R. Herrnstein, *The Matching Law: Papers in Psychology and Economics*, H. Rachlin and D. I. Laibson, Eds. Cambridge, MA: Harvard Univ. Press, 1997.

[31] R. Cogburn, "On products of random stochastic matrices," *Contemporary Math.*, vol. 50, pp. 199–213, 1986.

[32] E. Seneta, *Non-Negative Matrices and Markov Chains*, 2nd ed. New York: Springer-Verlag, 2006.

[33] M. Cao, A. S. Morse, and B. D. O. Anderson, "Reaching a consensus in a dynamically changing environment: Convergence rates, measurement delays and asynchronous events," *SIAM J. Control Optim.*, vol. 47, pp. 601–623, 2008.

ABOUT THE AUTHORS

Andrew Stewart (Member, IEEE) received the B.Sc. degree in mechanical engineering from the University of California San Diego, La Jolla, in 2006 and the M.A. and Ph.D. degrees in mechanical and aerospace engineering from Princeton University, Princeton, NJ, in 2008 and 2011, respectively.

He is now an Ocean Engineer in the Applied Physics Laboratory, University of Washington, Seattle. His interests include decision-making dynamics in mixed teams of humans and robots, dynamics and automated control of mechanical systems, and development of deployable hardware for exploration of remote environments. He has worked to develop a number of mobile sensor platforms, including unmanned air vehicles and autonomous underwater vehicles.

Ming Cao (Member, IEEE) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1999 and 2002, respectively, and the Ph.D. degree from Yale University, New Haven, CT, in 2007, all in electrical engineering.

From September 2007 to August 2008, he was a Postdoctoral Research Associate with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ. He is currently an Assistant Professor of Discrete Technology and Production Automation with the Faculty of Mathematics and Natural Sciences, University of Groningen, Groningen, The Netherlands. His main research interest is in autonomous agents and multiagent systems, mobile sensor networks, and complex networks.

Dr. Cao is an Associate Editor for *Systems and Control Letters*, and for the Conference Editorial Board of the IEEE Control Systems Society. He is also a member of the International Federation of Automatic Control (IFAC) Technical Committee on Networked Systems.

Andrea Nedic (Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Rutgers University, New Brunswick, NJ, in 2006 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 2011.

She has been exploring social decision making in association with the Princeton Neuroscience Institute, Princeton University. She is now a Senior Associate at Princeton Consultants in Princeton, NJ.

Damon Tomlin received the Ph.D. degree in neuroscience from Baylor College of Medicine, Houston, TX, in 2006.

He has been a Postdoctoral Associate in the Princeton Neuroscience Institute, Princeton, NJ. His research interests include reward-based decision making and social neuroscience.

Dr. Tomlin is a member of the Society for Neuroscience.

Naomi Ehrich Leonard (Fellow, IEEE) received the B.S.E. degree in mechanical engineering from Princeton University, Princeton, NJ, in 1985 and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1991 and 1994, respectively.

From 1985 to 1989, she worked as an Engineer in the electric power industry for MPR Associates, Inc. She is now the Edwin S. Wilsey Professor of Mechanical and Aerospace Engineering and an Associated Faculty Member of the Program in Applied and Computational Mathematics at Princeton University. She works in nonlinear control and dynamics with current interests in cooperative control, collective motion, and collective decision making. Applications include mobile sensor networks, collective animal behavior, and decision dynamics in mixed teams of humans and robots.