INVITED PAPER

Deterministic Modeling and Evaluation of Decision-Making Dynamics in Sequential Two-Alternative Forced Choice Tasks

lndividual decision making is studied using a deterministic decision-making model, and finite state machines for analysis of decision dynamics.

By Caleb Woodruff, Linh Vu, *Member IEEE*, Kristi A. Morgansen, *Senior Member IEEE*, and Damon Tomlin

ABSTRACT | The focus of the work in this paper is a systems-theoretic construction, analysis, and evaluation of a deterministic model of human decision making relative to experimental data. In sequential two-alternative forced choice decision tasks, a human subject is presented with two choices at every time step, is given finite time to select one of the choices, and receives a reward after a choice is made (presented as a number on a computer screen). The goal for the human is to obtain the maximal reward while not knowing the underlying reward assignment process. In this work, we present a parameterized deterministic model for human decision making in this context and analyze optimality and stability using a finite state machine approach. This model is then evaluated relative to experimental data from human subjects performing each of six tasks.

KEYWORDS | Finite state machines; human decision modeling; two-alternative forced choice tasks

Manuscript received October 12, 2010; revised September 23, 2011; accepted October 11, 2011. Date of publication December 22, 2011; date of current version February 17, 2012. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-07-10528.

C. Woodruff was with the Department of Aeronautics and Astronautics, University of Washington, Seattle, WA 98125 USA. He is now with Hood Technology Inc., Hood River, OR 97031 USA (e-mail: woodrc@uw.edu).

L. Vu is with the Department of International and Postgraduate Study, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam (e-mail: linhvu@ieee.org).

K. A. Morgansen is with the Department of Aeronautics and Astronautics, University of Washington, Seattle, WA 98125 USA (e-mail: morgansn@uw.edu).

D. Tomlin is with the Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544 USA (e-mail: dtomlin@princeton.edu).

Digital Object Identifier: 10.1109/JPROC.2011.2174105

I. INTRODUCTION

Recently, interest in human decision dynamics has been growing in the field of control systems with the goal of incorporating knowledge of human decision-making characteristics into the design of mixed teams of humans and robots to improve overall team performance [1]–[3]. For instance, if a design engineer knows that humans are more likely to understand and respond to computer outputs in a certain form, the engineer can make sure to incorporate that information into the design. Examples of potential applications of this line of research include mixed teams of humans and autonomous vehicles [4], [5], cooperation in construction tasks [6], and emergency response systems [7].

Relative to this framework, we focus our effort on studying one particular type of human decision making, namely *dynamic decision making*, a process in which a human makes a sequence of interdependent decisions in order to achieve some objective (for example, coordination of humans and robots in a foraging task [2]). The human receives feedback, also termed an *outcome* in this context, after every decision is made. The *environment*, which the human's decisions impact, changes as a result of the

human decisions, but it can also evolve automatically between decisions.

Dynamic decision making, in general, is a complicated process that is a function of both the particular task and the environment. To make the study of human decision making tractable, neuroscientists have devised simple dynamic decision-making experiments known as sequential two-alternative forced choice (TAFC) decision tasks [8] to study specific features of cognitive and psychological aspects of human decision making when the human has extremely limited information. In these tasks, a participant is presented with a binary choice and given a reward based on the current choice and the previous N choices. In human subject experiments, cognitive scientists have observed that for a majority of the human test subjects working with particular reward structures, decisions are centered around particular points, termed crossing points, where the reward return curves for the two options cross. For instance, in [9], this "matching behavior" was demonstrated in a task similar to the one we consider here. This particular bias indicates the existence of an underlying mechanism in human decision making in the sequential two-choice task.

Clearly, the mechanism of human decision making is an inherently stochastic process. Further, while in some instances decision making could be argued to be entirely stochastic, in many scenarios the process is composed of both a deterministic component and a stochastic component. Our belief is that the best models are those based on the most complete knowledge of both the deterministic and the stochastic elements as the more accurately a model is known, the better one can design around it. The intent in the work here is to explore the underlying structure of the deterministic component of the process in a particular context before stochastic components are incorporated. Extensions of the work here to include stochastic effects to the proposed underlying deterministic structure are the focus of ongoing research and will be discussed in the conclusion.

In order to mathematically investigate potential underlying mechanisms of human decision making, prior work has primarily addressed the use of Markov decision processes (MDPs) and drift-diffusion models (DDMs), a type of stochastic differential equation (SDE). MDPs apply to scenarios where events occur in discrete time and with discrete state values. Transitions from one state to another are treated probabilistically and are only dependent on the current state. SDEs, including DDMs, evolve in continuous time with continuous state values and are composed of both deterministic and stochastic elements. Relative to forced alternative tasks with two or more alternatives, a number of studies have utilized one or both of MDPs and SDEs to develop predictive models of decision making. Generally, DDMs have been used in studies of measurements of brain activity such as dopamine levels, neural firing rates, and functional magnetic resonance imaging (fMRI) imaging [8], [10]–[12] where data are collected during a period of time to determine the most likely selection of an action from a finite set at a decision time. One can then consider the set of actions as occurring at discrete times with continuous time data paths leading to each of the discrete actions. As mentioned above, these SDE models allow for both deterministic and stochastic elements and in work to date [13] have primarily involved simple linear or affine deterministic components that relate future decisions to weighting parameters determined from one or a few prior decisions. More complex and general frameworks of this type have been explored to a certain extent [14]–[16], but analytical complexity of SDEs with deterministic elements representing more complex reward-based decision-making strategies has generally been prohibitive. MDPs have been used to model decision making with strictly probabilistic transitions between decisions in a number of contexts [17], but more generally MDPs have been used in combination with DDMs. In particular, work in [18] considers TAFC tasks and a DDM together as a Markov process and shows that, with certain assumptions, the DDM will analytically exhibit matching behavior in similar scenarios as human subjects. In [19], convergence to a matching point is proven for a particular task termed the matching-shoulders-type task both for the model presented in this paper and for the DDM with a time decay extension termed the eligibility trace. In [12] and [20], a combination of DDM and MDP was used to address the empirical and analytical effects of social context (decisions and rewards of other people) on decision making.

The focus of the work here is to explore deterministic elements of human decision making. We consider a simple deterministic input–output model and investigate the limits of reasonable approximation of actual behavior. In our prior work [3], [21], control-theoretic tools were used to study, both analytically and experimentally, asymptotic behavior of human decision making in sequential two-choice tasks. In the work here, we extend our previous results and add a full analytic treatment of asymptotic behavior under the extended model from [21]. Further, the use of a tuning parameter is incorporated here to attempt to capture individual behaviors with regards to risk and reward. Work in [12] also addresses individual tuning, where their goodness-of-fit metric is the amount of time spent in binned locations, while our model is formulated to match decisions at each time step for purposes of identification of changes in human subject decision-making policies. Our deterministic decision model does successfully predict human decision making in such TAFC tasks much of the time (discussed below in detail). Compared to the prediction-error models [8], [10] for human decision making, our model is somewhat simpler but enables us to analyze asymptotic behaviors in sequential TAFC tasks for various types of reward structures. The results here have implications for directions to pursue in the use of combined MDP and SDE models of decision making.

The particular scenario considered here is, admittedly, abstracted from the goal of constructing methodologies for effective interaction of human operators and autonomous vehicles. However, the results of the analysis indicate approaches that can be taken in terms of data usage and presentation to human operators during certain types of reward-based activities. In particular, human subjects studies have shown that factors such as task switching [22], type of interface [22], [23], trust of automation [24], [25], and boredom and fatigue [26] can have either positive or negative effects on performance. The particular models and analyses here are intended to help identify decision-making characteristics such as tendencies to gravitate toward fixed points in reward, finding optimal rewards in the presence of local maxima, and changes in decision-making strategy. These characteristics may then be used to construct reward presentation interfaces that encourage operators toward certain desirable decision patterns, identify and respond to changes in decision-making strategies, and determine characteristics of beneficial operating environments.

The organization of the paper is as follows. In Section II, TAFC tasks are discussed relative to dynamics for the deterministic finite state machine approach. In Section III, we discuss the various reward structures used in the work here. We present our policy for decision making with mathematical analysis in Section IV, and comparisons of the predictions of the model to experimental data on human subjects are presented in Section V. In Section VI, we conclude and discuss future and ongoing work.

II. TWO-ALTERNATIVE FORCED CHOICE TASKS

The model and data presented here are all relative to the TAFC task. The TAFC is a simplified model that represents situations where a person must choose between fixed alternatives. Herrnstein, one of the early researchers in TAFC [9], used the example of a tennis game where the player must choose between a lob or a passing shot. If the player does nothing but lob the ball, the opponent will begin to expect that shot, resulting in the player losing points. Therefore, an optimal strategy must involve taking some combination of shots. In this section, the structure of the TAFC is outlined, and the dynamics that underly the problem are presented.

A sequential TAFC task is constructed as follows [8]. A participant is presented with two choices: A and B. The participant has to choose either A or B in a fixed amount of time, and after the decision is made, the participant receives a reward, presented as a number on the computer screen. The procedure is then repeated for a fixed number of iterations. The goal is to maximize the reward, which is calculated as follows. Let x be the percentage of A chosen among the last fixed number of N decisions. If the participant next chooses A, the reward is $\varphi_A(x)$; if the participant chooses B, the reward is $\varphi_B(x)$ (see Fig. 1).

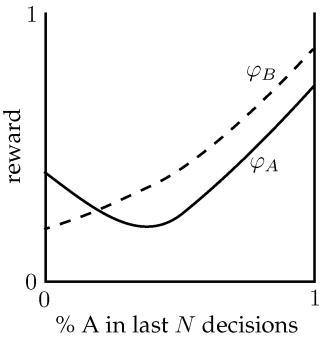


Fig. 1. *A typical reward structure in a sequential TAFC task. The solid line φ_A represents the reward for choosing option A. The dashed line φ_B represents the reward for choosing option B.*

The difficulty for a participant in the task is that the participant does not know how the reward is calculated. Yet, by exploring through a series of A and B choices and using the reward as feedback, a participant aims to achieve the maximum reward. Thus, this type of task is designed to explore human decision-making dynamics when the human has extremely limited information about the reward determination process.

Relative to a systems-theoretic context, TAFC tasks represent a class of optimization problems in dynamical systems that is different from classical control system problems in terms of the information available to the controller. Central in classical control systems theory of optimization is the assumption that the dynamics of the plant to be controlled are completely, or at least partially, known to the controller (such as in models with uncertainty or parameterized models with unknown parameters; in game theory, the rules of the game are known). In contrast, in TAFC tasks, the controller—the human in this case-knows nothing about the process, and the reward is available only after a decision is made.

We proceed by describing a deterministic mathematical model for reward dynamics in TAFC tasks (originally presented in [3]). Incorporation of decision policies will follow in the next section. Denote by $t_k \in [0, \infty)$, $k = 1, 2, \ldots$, the times at which the human makes decisions (also termed *decision times*) and by $u(t_k)$ the corresponding decisions, where $u(t_k) \in \{A, B\}$. Denote by the state x the percentage of A in the last N choices, where N is the window length (in the experiments below, $N = 20$). The dynamics of x are

$$x(t_k) =: \begin{cases} x(t_{k-1}) + \dfrac{1}{N}, & \text{if } u(t_k) = A, u(t_k) \neq u(t_{k-N}) \\ x(t_{k-1}) - \dfrac{1}{N}, & \text{if } u(t_k) = B, u(t_k) \neq u(t_{k-N}) \\ x(t_{k-1}), & \text{if } u(t_k) = u(t_{k-N}). \end{cases} \quad (1)$$

Note that $u(t_{k-N})$ is the choice that was most recently replaced by the moving window.

Denote by $\varphi_A : [0,1] \to \mathbb{R}$ a continuous function of the state x (called a *reward curve*) such that the reward will be $\varphi_A(x)$ when the percentage of A (including the current action) is x, and the human subject chooses A. Denote by $\varphi_B : [0,1] \to \mathbb{R}$ a reward curve such that the reward will be $\varphi_B(x)$ when the percentage of A is x, and the human subject chooses B. The output function for this process is defined to be

$$y(t_k) = \varphi_{u(t_k)}(x(t_k)). \tag{2}$$

We term the pair $\Gamma = (\varphi_A, \varphi_B)$ a *reward structure*.

The information available to the human operator at a decision time t_{k+1} is the set

$$\{u(t_1), \ldots, u(t_k), y(t_1), \ldots, y(t_k)\} =: I_{t_{k+1}}.$$

We include all past information in the set $I_{t_{k+1}}$, but humans do not generally have the ability to remember or to use all of the data from the start of the experiment through the current time k. The decision at time t_k (i.e., the control signal in a control systems context) is, in general, the output of a nonlinear map $\rho : I_{t_k} \to \{A, B\}$

$$u(t_k) = \rho(I_{t_k})$$

and, more generally, ρ could be the output of a dynamical system (i.e., the human decision-making dynamics) with input I_{t_k}. In the prediction-error model [8], such a ρ is a probabilistic map. We present here another view on human decision making in TAFC tasks by seeking a deterministic ρ, which is termed a *strategy* or a *policy*. As stated above, a deterministic ρ can be seen as an input–output simplification of the human decision-making dynamics, which enables us to more easily analyze asymptotic behaviors of the closed-loop system.

III. REWARD STRUCTURES

The particular set of reward-based tasks considered here were introduced in [27] and are shown in Fig. 2. Rather than represent specific real-world tasks, the different tasks are intended to isolate and to test for specific behaviors in humans. Specifically the converging Gaussian (CG) task is designed to push people towards the crossing point, while the behavior for maximal reward in the diverging Gaussian (DG) task is to stay at the crossing point, but by design it encourages subjects to explore. The rising optimum/reverse rising optimum (RO/RRO) and complex rising optimum/reverse complex rising optimum (CRO/RCRO) are pairs designed so that the optimum behaviors are difficult for a test subject to find but also so that simple decision patterns are encouraged [12].

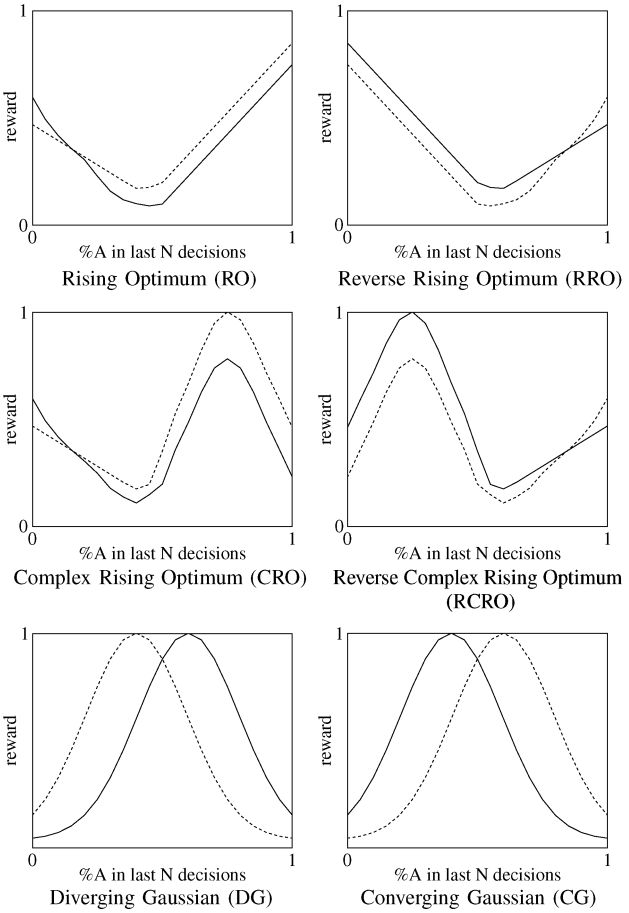


Fig. 2. ***Six types of reward structures for sequential TAFC tasks. The solid line is the reward curve for choice A, and the dashed line is for choice B.***

One aspect of decision making that will be considered here is whether the amount of change being experienced has any effect on human decision-making patterns. From the given reward structures shown in Fig. 2, the maximum possible change in reward for a single time step can be determined by simple differences. These values are reported in Table 1 for the reward structures in Fig. 2 and can be easily found for any reward structure. Note that the values are reported as absolute values and could be experienced by the subject as either an increase or decrease in

Table 1 Maximum Change in Reward for a Single Step With Step Size 1/20 (Can Occur as Positive or Negative Change)

Task	RO	RRO	CRO	RCRO	DG	CG
Max Change	0.1254	0.1254	0.3676	0.3676	0.4413	0.5228

one-step reward depending on the direction of change of the state x.

In the remainder of this section, we are concerned primarily with the mathematical characterization of different reward structures $\Gamma = (\varphi_A, \varphi_B)$. Specifically, we consider optimal decision paths, deconstruction of reward structures into basic subtypes, and asymptotic behavior over a composite reward structure.

A. Optimal Rewards

One can analytically construct the optimal average reward for each of the reward structures in Fig. 2. Let the average reward $\bar{y}$ be the value of the reward for a decision path that stays at exactly x_0 for all time, meaning that $u(t)$ is periodic of period N where N is the window length (e.g., $\underbrace{ABBABB\ldots ABB}_{N}ABB\ldots$). For any point x_0 along the x-axis, the average reward is given by

$$\bar{y}(x_0) = x_0\varphi_A(x_0) + (1-x_0)\varphi_B(x_0).$$

Example 1: Consider the reward structure shown in Fig. 3, which has the following functional description:

$$\varphi_A = \begin{cases} -0.6x + 0.34, & 0 \le x \le 0.4 \\ 0.8x - 0.22, & 0.4 < x \le 1 \end{cases}$$
$$\varphi_B = \begin{cases} -0.2x + 0.28, & 0 \le x \le 0.4 \\ 0.8x - 0.12, & 0.4 < x \le 1. \end{cases}$$

For any fixed $x = x_0$, the system in Fig. 3 has average reward

$$\bar{y} = \begin{cases} -0.4x_0^2 - 0.14x_0 + 0.28, & 0 \le x_0 \le 0.4 \\ 0.7x_0 - 0.12, & 0.4 < x_0 \le 1. \end{cases}$$

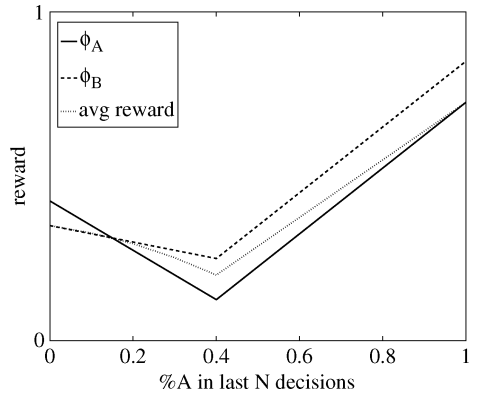


Fig. 3. *Example showing the average reward curve.*

Specifically, the periodic decision pattern

$$u(t_k) = \underbrace{AA\ldots AA}_{16\text{times}}\underbrace{B\ldots B}_{4\text{times}}$$

for which $x_0 = 0.8$, gives an average reward of $\bar{y} = 0.44$.

The best possible long-term reward is then to follow a decision path that maximizes $\bar{y}$. The particular value of x_0 that achieves this maximum can be reached in at most N steps, where N is the length of the history window. Using a similar construction, the maximum of each of the reward structures in Fig. 2 can be found as follows.

Theorem 1: For each of the reward structures in Fig. 2, the maximum possible average reward is given by

$$\bar{y}_{\max} = \max_{\bar{x}\in\{0,1/N,\ldots,1\}} \{\bar{x}\varphi_A(\bar{x}) + (1-\bar{x})\varphi_B(\bar{x})\}. \quad (3)$$

The decision policy to reach the optimal $\bar{y}$ is then to choose A $\bar{x}N$ times and B $(1-\bar{x})N$ times in a pattern that is periodic with period N, where $\bar{x}$ maximizes (3).

Proof: The result follows directly from the constructions above. ■

Note that while this result concerns an optimal return based on an average, the nonaverage optimal result will in fact be the same as the average result. To see this point, consider any of the tasks with state initialized at a value corresponding to the highest peak of the curves. The overall optimal reward corresponds to receiving that reward for all time. But receiving that reward requires that the state x, percent of A chosen out of the previous N decisions, remain constant. In some cases, this constant value can be achieved ($x = 0$ or $x = 1$ with the appropriate reward curve), but in general, a constant value of x requires a cyclic decision pattern.

Given the above theorem and the six reward structures of Fig. 2, inspection shows that the optimal reward for the RO task occurs at $\bar{x} = 1$ (a cyclic pattern of decisions being mostly A), the optimal reward for the RRO task occurs at $\bar{x} = 0$ (all decisions being A), the optimal reward for the CRO task occurs at the peak at approximately $\bar{x} = 0.8$, the optimal reward for the RCRO task occurs at the peak at approximately $\bar{x} = 0.2$, and the optimal reward for the DG and CG tasks occurs between $\bar{x} = 0.6$ and $\bar{x} = 0.4$.

B. Reward Structure Decomposition

Given an arbitrary reward structure, we would like to identify when and where a player following a given periodic, or cyclic, decision policy would receive rewards that demonstrate a periodic, or cyclic, pattern. When the periodic decision policy is such that the relative percentages of choices of each of the two options remain fixed

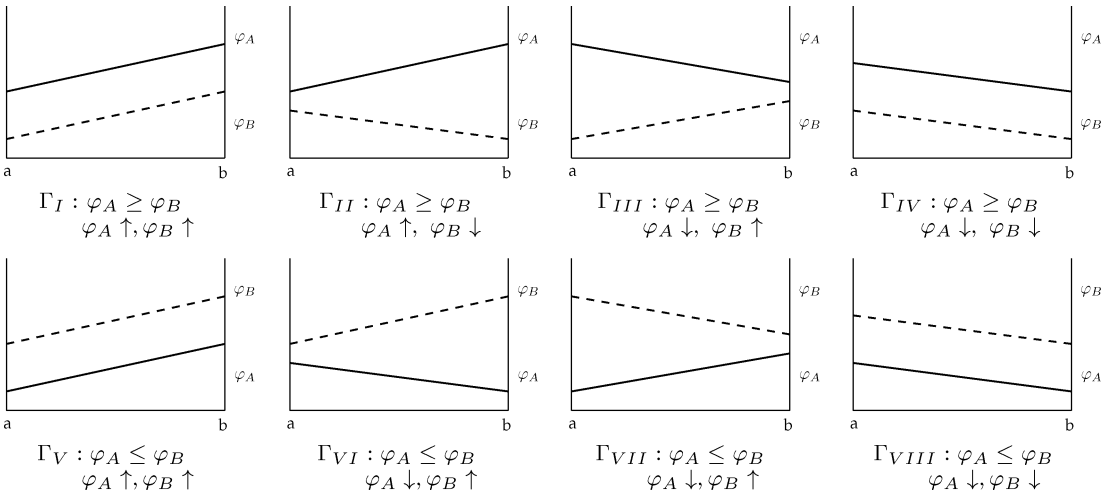


Fig. 4. *Basic reward structures for sequential two-choice tasks. The horizontal axis is the percentage of A in the last N decisions, and the vertical axis is the reward value for choice of A or B at a given percentage of A. The solid line φ_A is the reward for choosing A, and the dashed line φ_B is the reward for choosing B. Note: because the horizontal axis is %A, choosing A repeatedly moves the state x toward a value of b.*

and the resulting rewards are constant, the result is termed a "fixed point." Note that for fixed points, $x = x_0$ is constant. Relative to (1) such points occur for cyclic decision patterns of period N (the next decision is the same as the Nth previous decision). From an analytical perspective, these points and behaviors provide key information about what types of decision policies will lead to local versus global maximal rewards. Further, this information can be used to evaluate human decision-making tendencies under particular conditions and potentially alter those conditions to encourage desired human decision behaviors. In this section, we provide a tool to facilitate the fixed point analysis by showing how to decompose an arbitrary reward structure into a particular set of basis elements.

Analytical evaluation of fixed points and cyclic reward patterns in arbitrary reward structures can be accomplished via a deconstruction of the reward structure into a set of common basic components. Consider the eight basic reward structures shown in Fig. 4. By the notation $\varphi_A \geq \varphi_B$, we mean $\varphi_A(x) > \varphi_B(x) \forall x \in]a, b[$ with the endpoints characterized by $\varphi_A(a) \geq \varphi_B(a)$ and $\varphi_A(b) \geq \varphi_B(b)$. The definition of the basic reward structures in Fig. 4 is made such that $a, b \in [0, 1]$, and the structures are independent of the particular values of a and b. Further, note that the definitions of the basic types are independent of the vertical scaling. For mathematical purposes, this construction is convenient because φ_A and φ_B in each of the basic structures are monotonic.

Given these substructures, we can then state the following *substructure decomposition* lemma.

Lemma 1: Given an arbitrary reward structure $\Gamma = (\varphi_A, \varphi_B)$, if φ_A and φ_B are always either increasing or decreasing except possibly at finitely many points, then the structure Γ can be broken into a unique minimal finite string of substructures $\Gamma_1, \ldots, \Gamma_n$, where Γ_i, $i \in \{I, \ldots, VIII\}$, is one of the types in Fig. 4.

Proof: Let λ_A and λ_B, respectively, be the number of extremal points of the curves φ_A and φ_B, and let λ_{AB} be the number of crossing points of φ_A and φ_B. Let $x_i \in [0, 1]$, $i \in \{1, \ldots, \lambda_A + \lambda_B + \lambda_{AB} + 1\}$ be the abscissa values of each of the extremal and crossing points ordered such that $i < j \Rightarrow x_i \leq x_j$ (lexicographic ordering). Now, note that between any two successive extremal or crossing points (regardless of whether a given point is an extremal point for both curves or not), the curves φ_A and φ_B are either monotonically increasing or decreasing. As the basic reward structures include all possible combinations of monotonicity for two curves, the curves between any two extremal or crossing points must necessarily be one of the eight types $\Gamma_I, \ldots, \Gamma_{VIII}$. Clearly, then, an arbitrary reward structure $\Gamma = (\varphi_A, \varphi_B)$ can be expressed as the concatenation of substructures $\Gamma = \Gamma_1, \ldots, \Gamma_{\lambda_A + \lambda_B + \lambda_{AB} + 1}$. To see that this concatenation is uniquely minimal, note that removing any of the substructures would require that at least one remaining substructure has more than one extremal or crossing point in one or more of the curves which is not allowed. Further, note that breaking any substructure into a larger number of pieces would provide multiple substructures of the same type next to one another without additional extremal or crossing points. As the original structure could be decomposed using the smaller number of substructures, minimality follows from the original decomposition. Uniqueness is clear based on the categorization of the substructures. ■

Example 2: The CG reward structure in Fig. 2 has one extremal point in each of the two curves and one crossing

point. In order, these points are labeled x_1, x_2, x_3. The structure can then be uniquely and minimally decomposed into the sequence of four substructures Γ_I on $[0, x_1]$, Γ_{III} on $[x_1, x_2]$, Γ_{VI} on $[x_2, x_3]$ and Γ_{VIII} on $[x_3, 1]$.

IV. A POLICY FOR EXPLORATION AND EXPLOITATION

Given reward structures, we would like to find a deterministic strategy or policy that both accounts for finding desirable reward returns (e.g., fixed points, maximal reward) and captures the decision-making tendencies of humans presented with the given TAFC tasks. The mathematical analysis of a particular strategy will be presented in this section with a comparison to human subject results in the following sections.

A. The γ-Policy

In the psychology literature [28] and the game theory literature [29], a deterministic policy termed win–stay–loose–switch (WSLS) has been posed as being a model of human decision making in certain contexts. In this strategy, if a reward-based outcome occurs as expected based on a decision, i.e., when the last choice increases the reward, no incentive exists to change from using the last decision (alternatively, it is the incentive to continue with the same decision). If the outcome is not as expected, i.e., when the last choice decreases the reward, it is the incentive to switch the decision (in order to avoid further potential losses). Based on this deterministic decision-making strategy, we propose a simple decision-making rule, termed the γ-policy, for sequential TAFC tasks

$$u(t_k) := \begin{cases} u(t_{k-1}), & \text{if } y(t_{k-1}) \geq y(t_{k-2}) \\ \text{switch}(u(t_{k-1})), & \text{else} \end{cases} \quad (4)$$

where $\text{switch}(A) = B$, $\text{switch}(B) = A$, and $y(t_k)$ is given in (2). The slight difference between the γ-policy and [28] and [29] is that the reward can be the same instead of a clear win-or-lose situation. Additionally, only one player is present. We summarize this decision-making policy with the following two courses of action.

- A1) If the last reward is greater than or equal to the preceding reward, the previous decision choice is repeated.
- A2) If the last reward is less than the preceding reward, immediately switch to the alternative decision choice.

In cognitive psychology terminology, actions of type A1 are known as exploitive, and actions of type A2 are known as explorative (in fact, relationships between exploitation and exploration in human decision making are a major research theme in cognitive science [30]). Using this terminology, the first case of (4) encodes exploitation, and the second case encodes exploration. Switching between the two is triggered by whether reward expectations are met.

The actions A1 and A2 are idealistic. Some humans explore by switching decisions even if no decrease in rewards occurs (for example, in gambling types of activities), or some stay with a decision even if a temporary decline in rewards occurs (for example, such traits can be found in long-term investors).

Given these characteristics, we modify the γ-strategy to include a parameter that can be tuned to better capture human decision-making dynamics. The modified γ-strategy with a switching threshold δ is

$$u(t_k) := \begin{cases} u(t_{k-1}), & \text{if } y(t_{k-1}) \geq y(t_{k-2}) + \delta \\ \text{switch}(u(t_{k-1})), & \text{else.} \end{cases} \quad (5)$$

The decision remains the same only if the reward improves by at least δ. In essence, relative to the γ-strategy with behavior based on the reward obtained from the prior decision, the variable δ captures the risk attitude of humans: larger values of δ correspond to riskier behavior. One can also consider further generalizations of this extension of the γ-strategy to include additional parameters and dependence on decision and reward history beyond two steps. For purposes of brevity and clarity of presentation, such extensions are not considered here and are the subject of ongoing research.

B. Asymptotic Behavior for $\delta = 0$

Under the γ-policy, the decisions u can be shown to cause the system state to exhibit asymptotic behaviors. As in Section III-B with the basic reward structures, for generality, we allow the domain of the reward structure to be an arbitrary interval $[a, b]$ and use the state dynamics in (1) to explore the existence of fixed points and cyclic reward behaviors. An important point to note is that when considering reward structures with argument $[a, b]$ not necessarily equal to $[0, 1]$, the state x from (1) can limit to the edge of the interval $[a, b]$ rather than limiting to a fixed point within that interval.

Let $z(t_k) := (u(t_{k-1}), u(t_k))$ be the ordered sequence of two consecutive decisions at times t_{k-1} and t_k; clearly, $z(t_k) \in \{AA, AB, BA, BB\}$. We write $x \to x^*$ (respectively, $x \to \mathcal{S}$ where $\mathcal{S} = \{x^* - 1/N, x^*, x^* + 1/N\}$) if there exists $T < \infty$ such that $x(t) = x^*$ (respectively, $x(t) \to \mathcal{S}$) $\forall t \geq T$.

Lemma 2: The steady-state behavior of an agent following the γ-strategy in each of the basic reward structures is as follows. In all cases, if $x(t_1) = a$ and $z(t_1) = BB$, then $x \to a$; similarly, if $x(t_1) = b$ and $z(t_1) = AA$, then $x \to b$. Otherwise, we have the following.

- Γ_I: $x \to b$.
- Γ_{II}: If $z(t_1) = BB$, then $x \to a$; otherwise $x \to b$.

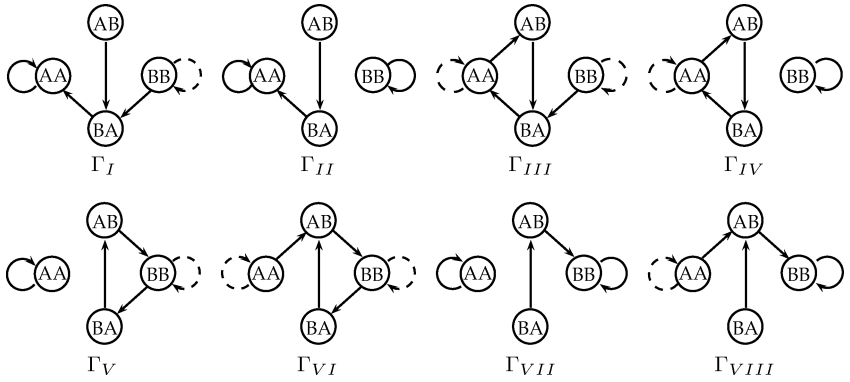


Fig. 5. *Finite state machines of decision dynamics for the basic reward structures in Fig. 4 with* $\delta = 0$. *Dashed lines are executed if* $y(t_{k-1}) = y(t_{k-2})$.

- Γ_{III}: If $a < 2/3 < 2/3 + 1/N < b$, then $x \to [2/3 - 1/N, 2/3 + 2/N]$; otherwise x limits to whichever of a or b is closer to 2/3.
- Γ_{IV}: If $z(t_1) = BB$, then $x \to a$; else if $a < 2/3 < 2/3 + 1/N < b$, then $x \to [2/3 - 1/N, 2/3 + 2/N]$; otherwise x limits to whichever of a or b is closer to 2/3.
- Γ_V: If $z(t_1) = AA$, then $x \to b$; else if $z(t_1) \neq AA$ and $a < 1/3 - 1/N < 1/3 < b$, then $x \to [1/3 - 2/N, 1/3 + 1/N]$; otherwise x limits to whichever of a or b is closer to 1/3 (mirror of Γ_{IV}).
- Γ_{VI}: If $b < 1/3 - 1/N < 1/3 < a$, then $x \to [1/3 - 2/N, 1/3 + 1/N]$; otherwise x limits to whichever of a or b is closer to 1/3 (mirror of Γ_{III}).
- Γ_{VII}: If $z(t_1) = AA$, then $x \to b$; otherwise $x \to a$ (mirror of Γ_{II}).
- Γ_{VIII}: If $x(t_1) = b$ and $z(t_1) = AA$, then $x \to b$; otherwise $x \to a$ (mirror of Γ_I).

Proof: Using the mathematical expression of the γ-policy in (4), construct a finite state machine for each of the eight basic reward structures of Fig. 4 (shown in Fig. 5). Note that four of the finite state machines have cycles of length three. The nature of the moving window history is such that if the next decision is the same as the decision N steps in the past, the value of the state x will not change; more succinctly, if $u(t_{k+1}) = u(t_{k-N})$, then $x(t_{k-1}) = x(t_k)$. This fact clearly implies that if $u(t_k) = u(t_{k-N-1})$ and $u(t_k) = u(t_{k-1})$, then $y(t_k) = y(t_{k-1})$, which under the γ-policy implies that $u(t_{k+1}) = u(t_k)$. Thus, any decision trajectory under the γ-policy that is periodic in N is necessarily a constant trajectory.

From the structure of the finite state machines, and a study of the initial conditions for special cases, the results of the lemma with regard to $\Gamma_I, \Gamma_{II}, \Gamma_{VII}, \Gamma_{VIII}$ clearly follow. For $\Gamma_{III} - \Gamma_{VI}$, the decision trajectory will necessarily be periodic, just not periodic in N (specifically $N = 20$, but the cycles are length three). Because the trajectory is not periodic in N, the state x will enter a limit cycle in the vicinity of 2/3 or 1/3 as appropriate. ■

Example 3: Consider the reward structure Γ_{III} under the γ-policy with $N = 7$, $a = 0$, and $b = 1$. Starting with choice history *BBBBBBA* at time t_0, the choice trajectory will be the result:

k	choice	history	x	reward
1	*A*	*BBBBBAA*	2/7	$y(t_1) < y(t_0)$
2	*B*	*BBBBAAB*	2/7	$y(t_2) < y(t_1)$
3	*A*	*BBBAABA*	3/7	$y(t_3) > y(t_2)$
4	*A*	*BBAABAA*	4/7	$y(t_4) < y(t_3)$
5	*B*	*BAABAAB*	4/7	$y(t_5) < y(t_4)$
6	*A*	*AABAABA*	5/7	$y(t_6) > y(t_5)$
7	*A*	*ABAABAA*	5/7	$y(t_7) = y(t_6)$
8	*A*	*BAABAAA*	5/7	$y(t_8) = y(t_7)$
9	*A*	*AABAAAA*	6/7	$y(t_9) < y(t_8)$
10	*B*	*ABAAAAB*	5/7	$y(t_{10}) < y(t_9)$
11	*A*	*BAAAABA*	5/7	$y(t_{11}) > y(t_{10})$
12	*A*	*AAAABAA*	6/7	$y(t_{12}) < y(t_{11})$
13	*B*	*AAABAAB*	5/7	$y(t_{13}) < y(t_{12})$
14	*A*	*AABAABA*	5/7	$y(t_{14}) > y(t_{13})$
15	go to $k = 6$			

As indicated by the finite state machine for Γ_{III} and confirmed in the chart, the system settles into a limit cycle of length three that lives in the set $x \in \{5/7, 6/7\} \subset [2/3 - 1/7, 2/3 + 2/7]$.

For reward structures that are not one of the eight basic types in Fig. 4, we can examine the behavior of the overall structure using Lemma 2 to obtain the following theorem.

Theorem 2: Consider the system (1) with rewards (2) under the γ-policy (4) with a reward structure Γ on $[0, 1]$,

such that Γ meets the requirements of Lemma 1. For any initial values of $u(t_{k-N}), \ldots, u(t_{k-1})$, there is a time $T < \infty$ and a value $x^*(u(t_{k-N}), \ldots, u(t_k)) \in [0,1]$ such that $|x(t) - x^*| \leq 1/N$ for all $t > T$.

Proof: Given that the reward structure Γ meets the requirements of Lemma 1, Γ can be broken into a unique string of substructures from Fig. 4. The results of Lemma 2 show that within a single substructure, the behavior of an agent under the γ-policy will tend toward a fixed point. At the junction of two adjacent basic reward structures, the junctions will either be locally stable or unstable. In the case of a locally stable junction, where both substructures tend to push the agent toward the junction (sources), the agent will oscillate around the junction or stop at the junction. In the unstable case, two subcases exist. In the first subcase, one substructure is a source and the other pushes the agent away from the junction (a sink). In this case, the agent will travel from the direction of the source toward the sink. In the second unstable subcase, two sinks meet at the junction. In this case, the agent will stay in the substructure in which it starts and behave according to the local substructure behavior from Lemma 2. The agent will then follow the behavior in Lemma 2 until it reaches either a stable equilibrium or a fixed point within a substructure as defined in Lemma 2. ■

Example 4: Consider again the decomposition of the CG reward structure presented in Example 2 and an agent following the γ-policy. From the edge conditions, if the agent starts at state $x = 0$ with $z = BB$, the state will stay at $x = 0$. Similarly, if the agent starts at state $x = 1$ with $z = AA$, the state will stay at $x = 1$. Otherwise, we have the following.

- If the state is in segment one Γ_I, the state will travel through x_1 into segment two.
- If the state is in segment two Γ_{III} and if $x_1 > 2/3$, the state will travel to the left and enter a limit cycle around x_1. If $x_1 < 2/3 < x_2$, the state will enter a limit cycle around 2/3. Otherwise, the state will travel through x_2 into the next segment, potentially entering a limit cycle around x_2.
- If the state is in segment three Γ_{VI} and if $x_2 > 1/3$, the state will travel left into segment two. If $x_2 < 1/3 < x_3$, the state will enter a limit cycle around 1/3. Otherwise, the state will travel through x_3 and enter a limit cycle around x_3.
- If the state is in segment four Γ_{VIII}, the state will travel left through x_3 into segment three.

Depending on initial location and the values of x_1, x_2, x_3, five potential locations exist around which the agent's state can limit cycle: $x_1, x_2, x_3, 1/3, 2/3$.

C. Asymptotic Behavior for $\delta \neq 0$

In the original γ-policy, when $\delta = 0$, three factors determine system behavior (see Fig. 6).

- Is $\varphi_A > \varphi_B$?

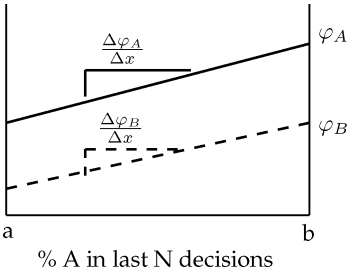


Fig. 6. *Features of reward structure for analysis with $\delta \neq 0$.*

- Is $(\Delta\varphi_A/\Delta x) \geq 0$? (Continue using $u = A$ in the γ-policy as x is increasing.)
- Is $(\Delta\varphi_B/\Delta x) \leq 0$? (Continue using $u = B$ as x is decreasing.)

We define $\Delta\varphi_A/\Delta x$ to be the change in reward as the player moves one step to the right along the reward curve φ_A; similarly for $\Delta\varphi_B/\Delta x$. From these conditional tests, the eight (or 2^3) substructures in Fig. 4 are derived. Similarly, when $\delta \neq 0$, the following conditions determine system behavior.

- Is $\varphi_A > \varphi_B$?
- If $\delta > 0$, is $\varphi_A - \varphi_B \geq \delta$? Otherwise, if $\delta < 0$, is $\varphi_B - \varphi_A \geq \delta$?
- Is $(\Delta\varphi_A/\Delta x) - \delta \geq 0$?
- Is $(\Delta\varphi_B/\Delta x) + \delta \leq 0$?

These inequalities lead to 32 possible conditions; here we consider only cases where $\varphi_A > \varphi_B$ because in the opposite case behavior will be mirrored. Proceeding with $\varphi_A > \varphi_B$, we are left with 16 possible scenarios which are listed in Table 2. Each column represents a possible combination of the relevant conditions where "✓" means the particular condition is true, and "x" means the condition is false [e.g., an x in the second row means that for the particular combination of conditions, $(\Delta\varphi_A/\Delta x) \not\geq 0$].

Behavior under the modified γ-policy with $\varphi_A > \varphi_B$ is as follows.

- From our initial assumption, $\varphi_A - \varphi_B > 0$, so if $\varphi_A - \varphi_B < |\delta|$, then we have the following.
  - For $\delta > 0$, a switch from A to B will decrease the reward and trigger a switch. The gain in reward from switching back to A is not greater than δ, however, and the agent will immedi-

Table 2 Possible Conditions on Slope of Reward Structure Curves and Value of δ; See Fig. 6. Each Column Represents a Full Set of Conditions Where "✓" Is True and x Is False. For $\delta < 0$, See Fig. 7 for Finite State Machines. For $\delta > 0$, See Fig. 8

$\varphi_A - \varphi_B < \lvert\delta\rvert$	✓	x	✓	✓	✓	x	x	x
$\frac{\Delta\varphi_A}{\Delta x} - \delta \geq 0$	✓	✓	x	✓	x	✓	x	x
$\frac{\Delta\varphi_B}{\Delta x} + \delta \leq 0$	✓	✓	✓	x	x	x	✓	x
FSM	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)

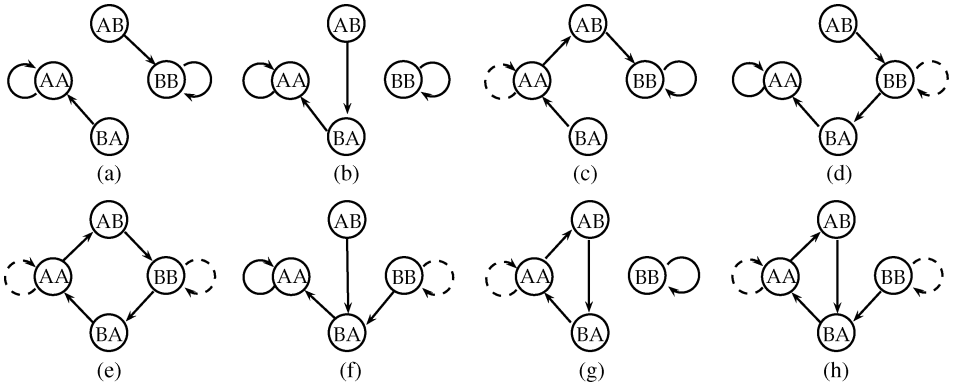


Fig. 7. *Representative finite state machines for $\delta < 0$. Dashed lines are executed if $y(t_{k-1}) = y(t_{k-2})$.*

ately switch back to A, causing continuous cycling.

- For $\delta < 0$, a switch from A to B will not decrease rewards enough to trigger a switch back to A, so the agent will stay with B for the next choice; similarly, a switch from B to A will increase the reward, and the agent will use A for the next choice.

- If $(\Delta\varphi_A/\Delta x) - \delta \geq 0$, a step in the positive x direction along φ_A causes a change in reward that is more positive than δ, so the agent will stay with A for the next decision.
- If $(\Delta\varphi_B/\Delta x) + \delta \geq 0$, a step in the negative x direction along φ_B causes a change in reward that is more positive than δ, so the agent will stay with B for the next decision.

The finite state machines that correspond to $\delta < 0$ are given in Fig. 7, and those for $\delta > 0$ are given in Fig. 8. With these finite state machines, we can state the following result for asymptotic system behavior.

Lemma 3: The steady-state behavior of an agent following the modified γ-strategy with $\delta < 0$ in each of the reward structures outlined in Table 2 is as follows. In all cases, if $x(t_1) = a$ and $z(t_1) = BB$, then $x \to a$. Similarly, if $x(t_1) = b$ and $z(t_1) = AA$, then $x \to b$. Otherwise, we have the following.

- Γ_a: If $z \in \{AA, BA\}$, then $x \to b$; otherwise $x \to a$.
- Γ_b: If $z = BB$, then $x \to a$; otherwise $x \to b$.
- Γ_c: $x \to a$.
- Γ_d: $x \to b$.
- Γ_e: If $a < 1/2 < b$, then $x \to 1/2$; otherwise x limits to whichever of a or b is closer to 1/2.
- Γ_f: $x \to b$.
- Γ_g: If $z = BB$, then $x \to a$. If $a < 2/3 < b$ and $z \neq BB$, then $x \to [2/3 - 1/N, 2/3 + 2/N]$; otherwise x limits to whichever of a or b is closer to 2/3.
- Γ_h: If $a < 2/3 < b$, then $x \to [2/3 - 1/N, 2/3 + 2/N]$; otherwise x limits to whichever of a or b is closer to 2/3.

Proof: Following the proof of Lemma 2, construct finite state machines according to the modified γ-policy in (5) for each possible reward structure configuration in Table 2 (Fig. 7). The results follow. ■

As will be noted below, data show that most people do not have a best fit δ value above zero. In the interest of

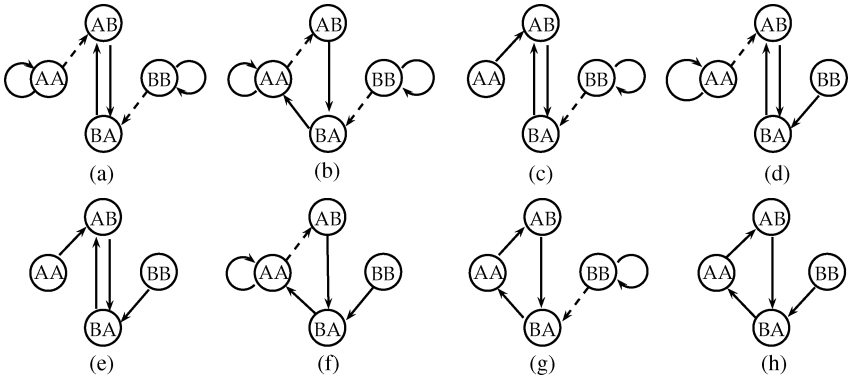


Fig. 8. *Finite state machines for $\delta > 0$. Dashed lines are executed if $y(t_{k-1}) = y(t_{k-2})$.*

brevity, we omit the analytical treatment of the $\delta > 0$ case, although we note that an agent following the extended γ-policy with positive δ will generally end up with $x \in \{1/3, 1/2, 2/3\}$.

V. MODEL EVALUATION

Given the proposed decision policy and analysis of the mathematical properties of the policy, we would like to determine how well human decision making corresponds to our theoretical deterministic decision policy. To this end, our theoretical results have been compared to human subject results. In particular, we discuss below instantaneous human decision making relative to the γ-policy, selection of appropriate switching threshold δ and effects of that threshold on the matching results, and qualitative tendencies of human subjects to gravitate toward either fixed points or maximal reward.

A. Experimental Setup

The data used here comes from experiments carried out at Princeton University and at Baylor University [27]. The subjects participated in sequential TAFC tasks with each task corresponding to one of the six types of reward structures shown in Fig. 2. In the data set used for evaluation here, 60 participants played using each of the six reward structures. For each task, the subject was allowed to make a finite number of choices (either 220 or 150 choices), and a single task experiment lasted about 35 min (in the case of 220 choices). For consistency in analysis of the data, only the first 150 choices were used from any of the experiments. In order to offset any learning effects in the experiment, subjects only performed each task type once, and tasks were presented to a subject in a random order with 10 s between two different task experiments. Because reward calculation requires 20 past choices, the first 20 choices were prechosen, but they were the same for all subjects in the same task.

B. Average Behavior

To qualitatively compare our analysis of fixed points and optimal average reward to actual human behavior, we considered the behavior of the entire group of subjects. Using the analysis tools from Theorem 2 shows that each task has a fixed point at $x = 0$ and $x = 1$, the DG has additional fixed points at $x = 1/3$ and $x = 2/3$, and the CG has a fixed point at $x = 1/2$. In Fig. 9, the average behaviors are shown for the group of 60 people that completed the tasks in the alone condition. The full length vertical lines are at the fixed points listed above. Clearly, in all tasks other than the DG, the subjects tended to spend most of their time at fixed points. Interestingly, the only case in which subjects tended toward the global optimal reward policy was in the CG task. In the cases of the RO, RRO, CRO, and RCRO tasks, subjects gravitated toward locally optimal reward policies but did not explore enough to find the global reward policy. In the DG task, no strong tendency toward a particular policy emerged. Note that in each case, some number of subjects did find a policy that led to the optimal reward, but such cases were not the general rule.

To determine how the γ-strategy would independently perform with the same initial conditions as the human

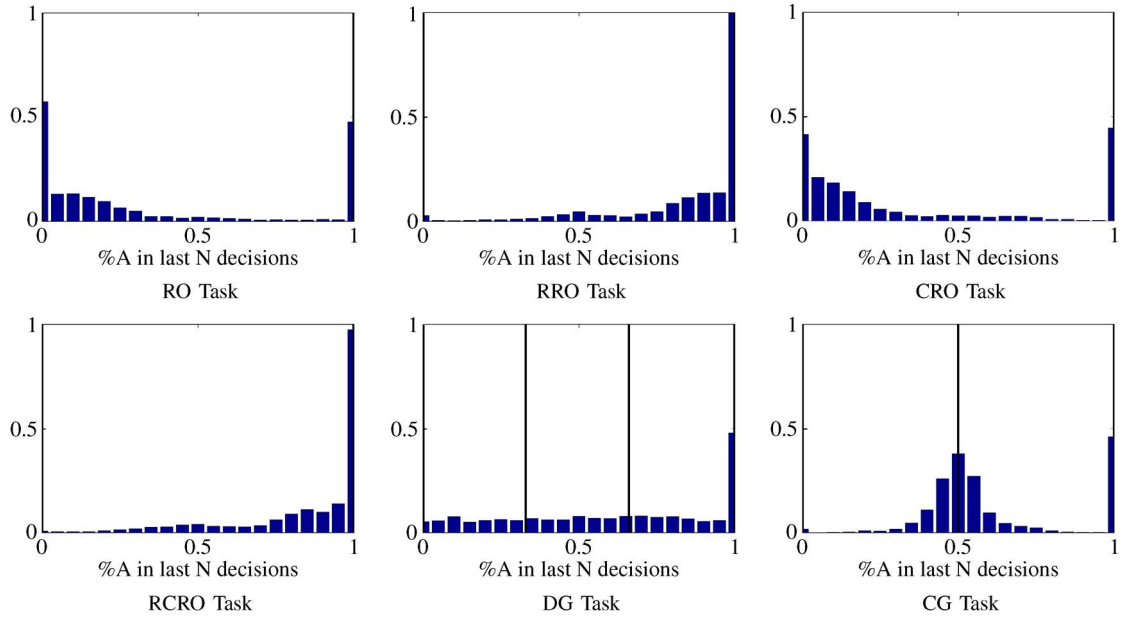


Fig. 9. *Normalized total behavior under each reward structure in Fig. 2. Full length vertical lines in DG and CG tasks represent fixed points from analysis under Theorem 2. All tasks have fixed points at zero and one.*

subjects, we can use the analysis from Theorems 1 and 2 with the initial conditions from the human subject tests. The initial conditions for the four tasks are $x_0 = 0, 1, 0, 1, 0.4,$ and 0.75. Based on these initial conditions, the γ-strategy would lead to the respective fixed points of $x_f = 0, 1, 0, 1, 0.66,$ and 0.5. These results are consistent with the human subject results and correlate with the fact that, because decisions in the γ-strategy are only based on the two most recent rewards and decisions, the overall results are generally only locally optimal. Reaching the global optimum requires starting at appropriate initial conditions.

C. Prediction Methodology

To evaluate our model relative to experimental data, we applied the rewards from the experiments to the model and then compared the actual human decisions with those predicted by the model. In this way, we directly compared the prediction of our model with the actual human decision at every time step during the experiments. Specifically, let $\{\tilde{u}_i^j(t_k)\}$ and $\{\tilde{y}_i^j(t_k)\}$ be the experimental decisions and the experimental rewards for participant i with reward structure j, where $i = 1, \ldots, \mathcal{N}$, with $\mathcal{N}$ the number of subjects (here, $\mathcal{N} = 60$) and $j = 1, \ldots, 6$ (for enumeration purposes, task 1 is the RO, task 2 is the RRO, task 3 is the CRO, task 4 is the RCRO, task 5 is the DG, and task 6 is the CG). The predicted decision using experimental data $\hat{u}_i^j(t_k)$ was defined as

$$\hat{u}_i^j(t_k) := \begin{cases} \tilde{u}_i^j(t_{k-1}), & \text{if } \tilde{y}_i^j(t_{k-1}) \geq \tilde{y}_i^j(t_{k-2}) + \delta \\ \text{switch}\left(\tilde{u}_i^j(t_{k-1})\right), & \text{else.} \end{cases}$$

Define the match variable $m_i^j(t_k)$ for participant i with reward structure j at decision step t_k as

$$m_i^j(t_k) = \begin{cases} 1, & \text{if } \hat{u}_i^j(t_k) = \tilde{u}_i^j(t_k) \\ 0, & \text{else.} \end{cases}$$

Note that a value of 1 corresponds to a match between theory and experiment, and a value of 0 corresponds to a mismatch. For participant i in task j, define the percentage of matching of predictions by the model as

$$p_i^j := \frac{1}{k} \sum_k m_i^j(t_k) \tag{6}$$

and let

$$n^j := \left(\text{card}\left\{i | p_i^j > P/100,\ i = \{1, \ldots, \mathcal{N}\}\right\}\right)/\mathcal{N}$$

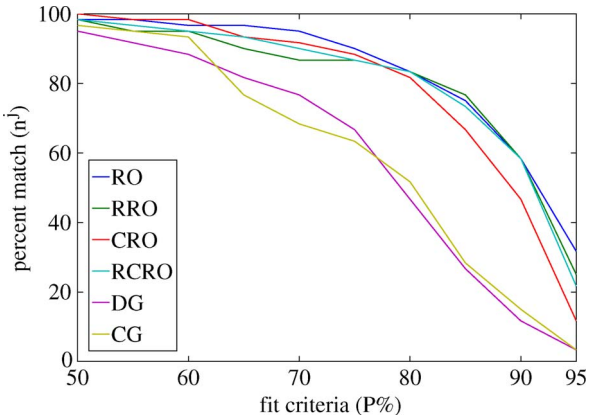


Fig. 10. *Percentage of matching between γ-strategy and experimental data with $\delta = 0$ for each of the tasks at a range of matching levels P%.*

be the percentage of the number of subjects i such that $p_i^j > P/100$ (i.e., the number of subjects having more than P% matches with the model).

D. Results for $\delta = 0$

The results for the γ-strategy with $\delta = 0$ are shown in Fig. 10 for a range of values of P. The lower end of the plot, with $P = 50$, corresponds to how often the γ-strategy performed at least as well as the completely random strategy of flipping a coin. The upper end of the plot corresponds to how often the γ-strategy performed exactly the same as a human subject. At a fit percentage of 50%, the data indicate that, as hoped, the γ-strategy indeed performed at least as well as random choice. The continued large matching rates as the fit percentage was increased, up to the range of 60%–70% for most of the tasks, provides strong support for the hypothesis that decision making in these TAFC tasks possesses a definitive deterministic element and that the γ-strategy is a reasonable model of that element. Beyond the level of 70% fit percentage, the matching rate falls off more or less quickly depending on the task suggesting that perhaps different strategies were being used by different human subjects or different strategies were being used during a single experiment by a single subject. This difference in strategy may correspond in some cases to alternating between a fixed decision policy and one that is random, or some combination of the two. The incorporation of appropriate switching between strategies and some of these strategies being stochastic is the subject of ongoing and future work.

The selection of a particular value of P to indicate a "good" fit of model to data is somewhat subjective. As a reasonable compromise between "better than completely random" and "completely matched in all cases," we will utilize a value of $P = 85$ in the remaining discussion. The results for $P = 85$ are tabulated in Table 3. An observation from Table 3 is that the γ-strategy correctly captured

Table 3 Model Validation for the γ-Strategy in Each of the Six Tasks in Fig. 2 With Matching Level at $P = 85$

Task	RO	RRO	CRO	RCRO	DG	CG
n^j	0.68	0.65	0.62	0.63	0.10	0.17

human decision making at a level of at least 85% in the scenario here for about two thirds of the human test subjects except for the Gaussian tasks. From the disparity in results, clearly the underlying structure has a large effect on how strongly people followed a WSLS-type strategy. For instance, the DG had a very low percentage of match at 10%, while the RO, at 68%, apparently encouraged WSLS decision making. Interestingly in [12], the average likelihood values used to determine goodness of fit were much higher for the simple and complex rising optimum tasks (they considered the task and reverse together) than for the Gaussian tasks. Our results are completely in line with theirs in this regard.

E. Results for $\delta \neq 0$

For the modified model with threshold (5), we used the same set of data as above and fit values of δ in three scenarios: 1) one value of δ for all participants for each task; 2) one value of δ for each participant in all tasks; and 3) one value of δ for each participant in each task. As the data show below, while the values of δ demonstrated trends, they were not constant for the different tasks, for a given user, or even during a single experiment.

The value of δ was chosen to be the smallest $|\delta|$ that maximized either the percentage number of participants who have more than $P\%$ matches between model prediction and actual data (scenario 1) or the number of fit choices (scenarios 2 and 3). Note that as before, we take $P = 85$ as shown in Fig. 10 where $P = 85$ is a reasonable compromise between number of people fit and fit accuracy. In scenario 1, some dependence on P was present because we chose δ such that n^j was maximized subject to the value of P. In the other scenarios, δ did not depend on P because δ was chosen as the absolute best fit for each person which, in the end, did not depend on P.

An important point to recall from Section III is that the possible change in reward in a single time for any state x has a maximum value. For instance, if the policy in (5) is applied with a value of δ that is more negative than the greatest possible decrease in rewards, a switch will never be predicted, and the model effectively becomes $u(t_k) = u(t_{k-1})$. Thus, to determine δ, we fit the largest number of matched choices and minimize $|\delta|$ to prevent the value of δ from being too positive or too negative relative to the reward structure.

1) Determination of a Single δ for All Participants for Each Task: In this case, we found δ^j across all participants for each task j, $j = 1, \ldots, 6$. Such a value of δ for each task should give us an indication of the tendency of a given reward structure to promote exploitive (or alternatively, explorative) tendencies of participants. The value of δ^j was found as follows:

$$\delta^j = \arg\min_{\delta \in [-1,1]} \left(|\delta| - n^j(\delta) \right)$$

which means finding the minimal $|\delta|$ that maximized the number of participants with greater than 85% match (n^j is always positive so $\max n^j = -\min n^j$). Generally multiple values of δ can be found, so the one of smallest magnitude was chosen. In the results here, we numerically found δ^j using a grid search with quantization level 0.01 in the range $[-1, 1]$. The results are tabulated in Table 4. Compared to the results with no threshold, the matching improved significantly (percentage improvement indicated by $\%\Delta$) ranging between 5% and 22% improvement. The values of δ found imply that the RCRO task leads to less exploration in the decision space than the other tasks (δ is more negative), and that decision making in the CRO task is most explorative.

Table 4 Model Validation for the γ-Strategy With Single δ for All Humans in a Given Task

Task	RO	RRO	CRO	RCRO	DG	CG
n^j	0.73	0.75	0.68	0.67	0.32	0.23
δ^j	-0.01	-0.08	-0.06	-0.03	-0.12	-0.08
$\%\Delta$	5%	10%	9%	4%	22%	6%

While the number of participants having more than 85% matched predictions increased in this case compared to the case without the threshold δ, the percent matching in the Gaussian tasks was still only 32% and 23%, respectively. However, human decision making can be quite diverse, and even in scenarios as simple as the TAFC task, the number of possible strategies is astronomical—for n binary choices, 2^n unique decision paths are possible. Yet about 70% of the participants followed the γ-strategy in the RO tasks more than 85% of the time.

2) Determination of a δ Value for Each Participant Across All Tasks: To determine whether the risk tendencies of human subjects are fairly consistent or vary between subjects, we calculated the value of δ for each subject, $i = 1, \ldots, \mathcal{N}$, in all tasks such that

$$\delta^i = \arg\min_{\delta \in [-1,1]} \left(|\delta| - \sum_{j=1}^{6} p_i^j(\delta) \right).$$

Again, we numerically found δ^i using a grid search with quantization level 0.01 in the range $[-1, 1]$, and p_i^j is given

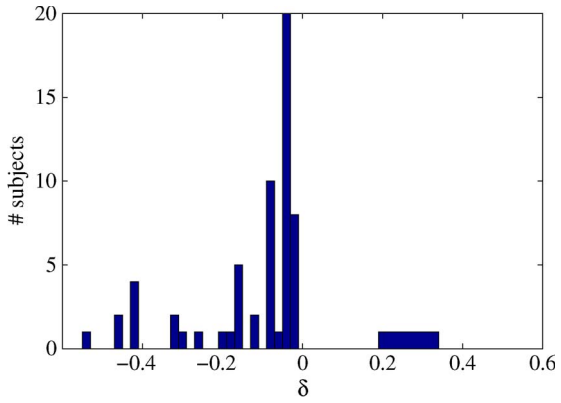


Fig. 11. ***Distribution of δ^i for each participant across all six tasks.***

Table 5 Model Validation for γ-Strategy With a Single δ for Each Human Across All Tasks

Task	RO	RRO	CRO	RCRO	DG	CG
n^j	0.75	0.77	0.67	0.73	0.27	0.28
%Δ	7%	12%	5%	11%	17%	11%

in (6). The distribution of δ^i among the 60 test subjects is plotted in Fig. 11. The values for the majority of the subjects are fairly distributed between 0 and −0.2. Also, almost all of the participants had a negative δ (and thus, are conservative in the decision space), and only three out of the 60 subjects had a positive or explorative δ. Table 5 shows only a small improvement between this case and that in Section V-E1 and shows a decline in the DG task.

3) Determination of a δ Value for Each Participant in Each Task: We next calculated the smallest $|\delta_i^j|$ for each participant i in each task j that maximized the percentage of matched predictions for the participant. The calculation of δ_i^j for each human in each task was made such that

$$\delta_i^j = \arg\min_{\delta\in[-1,1]} \left(|\delta| - p_i^j(\delta) \right) \tag{7}$$

again using a grid search with quantization level 0.01 in the range $[-1, 1]$. The histograms of these results are plotted in Fig. 12, and the means and variances of the values of δ are given in Table 6. One observation is that values for δ were quite concentrated for the RO and RRO tasks, with much smaller variances compared to other tasks. This result makes intuitive sense because the RO and the RRO were designed to encourage simple choice patterns, either all *A* or all *B*. The comparisons of the results to the original γ-strategy are indicated by %Δ. These results both improved relative to the original strategy and relative to the use of a single value of δ for each task. As one would expect, tuning the parameter δ to each subject led to improved performance of the system. Here, we see that our results have improved for the Gaussian tasks but are still far short of the fit level of the

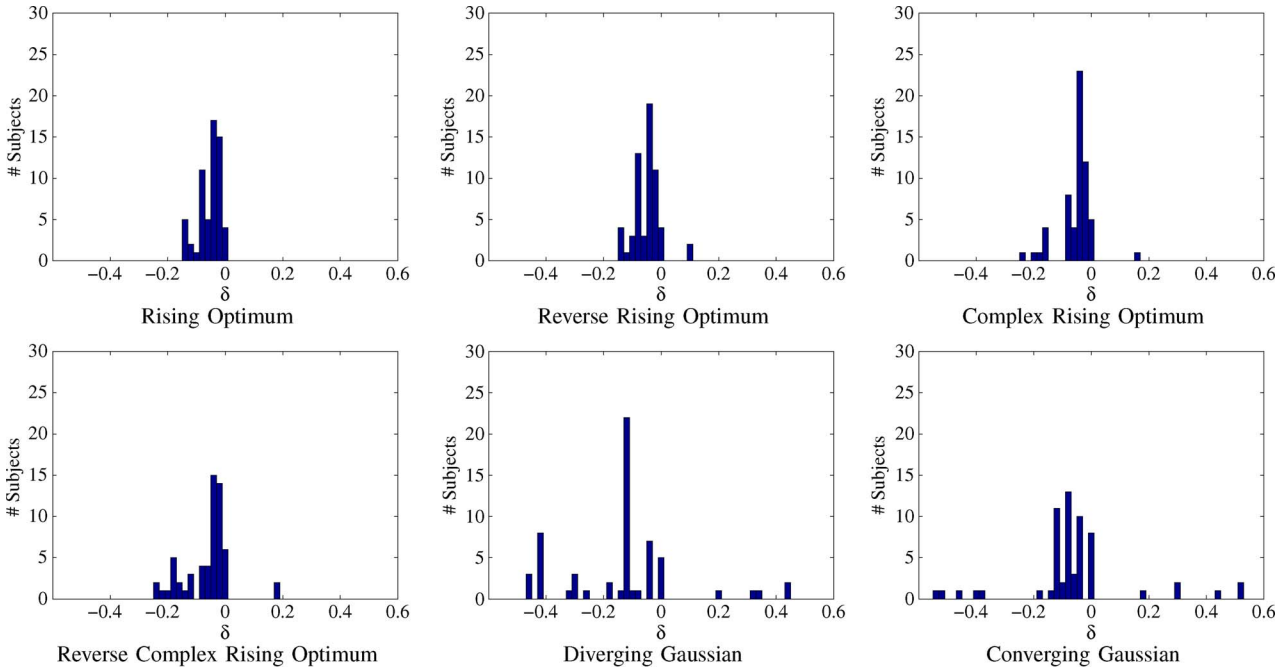


Fig. 12. ***Distribution of δ_i^j for all participants in each task j.***

Table 6 Mean and Variance of δ Calculated for Each Human in Each of the Six Tasks. Comparison to Original γ-Strategy Result With $\delta = 0$ Is Given by $\%\Delta$

Task	RO	RRO	CRO	RCRO	DG	CG
n^j	0.78	0.78	0.73	0.75	0.37	0.37
$E(\delta)$	-0.05	-0.04	-0.05	-0.06	-0.13	-0.05
$Var(\delta)$	0.0014	0.0021	0.0035	0.0067	0.0415	0.0363
$\%\Delta$	10%	13%	12%	18%	27%	20%

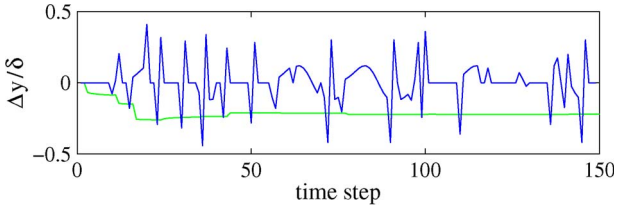


Fig. 13. *Plot of $\triangle y$ at each time step with the best fit δ for each time step shown.*

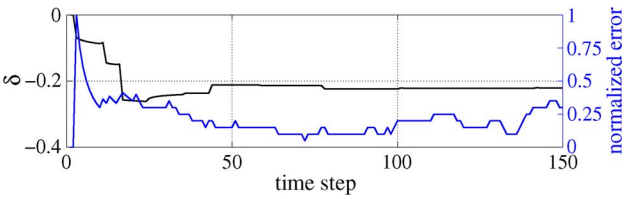


Fig. 14. *Best fit δ at each time step along with normalized error over last twenty time steps.*

other tasks. Again, these results are in line with the results from [12].

4) Adaptive Decision Strategy: The possibility exists that the human subjects change their decision-making strategy over the course of the experiment. Such changes could be the result of learning, boredom, or many other possibilities. One approach to identifying whether such changes occurred is to calculate δ at each time step for a given subject. If such a quantity does allow for the identification of changes in decision making, one could consider running the identification in real time with the goal of providing alternative cues to change subject behavior.

To find the best fit values of $\delta(t_k)$ at each time step for a given set of task data, we ran through the grid search described in Section V-E1 for each time step t_k. This time history allowed us to see stepwise changes in the value of δ (e.g., Fig. 13). The value of δ at a given time step is not in general the largest change in reward the subject has seen, but instead it is the best compromise between large losses when the subject does not change decisions and smaller losses where the decision is changed. A representative case of the results from this approach is shown in Figs. 13 and 14. The best fit δ compared to the change in reward for each time step is shown in Fig. 13, while the best fit δ along with the errors in the last $N = 20$ time steps is shown in Fig. 14. Clearly, when errors occur, the best fit value of δ is changed in order for our model to accommodate the new strategy choices. For example, as shown in Fig. 14, as our prediction mechanism starts to make errors around time step 75, the parameter value is updated slightly, and again at time step 100. This result shows that the parameter is responding to missed predictions which may indicate the subject temporarily switching to a more complicated choice pattern.

5) Difficulties in δ Fitting: We determined the best fit value of δ in all cases by evaluating over a grid, finding the values of δ that gave the highest percentage match, and then selecting from that set the value of δ with minimal magnitude. As can be seen in Fig. 15, this method resulted in values of δ that were closely related to the changes in reward the person actually saw. For instance, in Fig. 15(c), the best fit δ was the next discrete step below the largest decrease in reward for a time step with the result that our model predicted no switches. In other cases, the best fit value may be meaningless. As shown in Fig. 16, often no

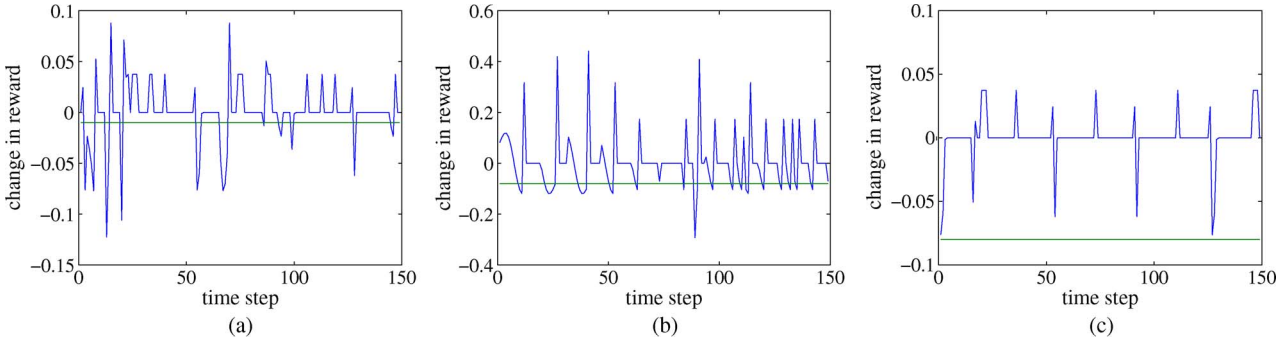


Fig. 15. *Change in reward at time step compared to the best fit δ. (a) Subject 1, task 1. (b) Subject 1, task 4. (c) Subject 3, task 1.*

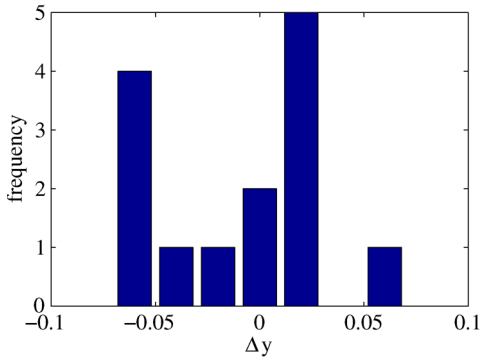


Fig. 16. ***Subject 3, task 1, change in reward before a switch.***

discernible relationship existed between change in reward and switching decisions. This lack of relationship makes predicting switches based on information about change in reward unlikely with the γ-strategy.

VI. CONCLUSION

In this paper, we have considered the use of a deterministic reward-based decision-making policy with parameterized threshold as a model of human decision making in TAFC tasks. The capabilities of the model relative to these tasks was considered both from an analytical point of view and in comparison to results from human subjects testing. The model system and the human subjects both demonstrated consistent averaged results in tending toward fixed points of the reward curves but generally did not gravitate toward global optima. In direct comparison of decision-by-decision results, the results with and without a parameterized bias in the response to a decrease in reward showed levels of correlation in many cases that indicated at least some underlying use of the WSLS γ-strategy as a decision-making policy. Specifically, in the case of no bias, results in four of the six tasks had a match between theory and experiment at a level over 85% for more than 60% of the human subjects. For the other two tasks, the matching level was only in the range of 10%. Note, however, that the rate of matching become significantly higher with thresholds lower than 85% but still above 50% where no correlation is present between predicted and actual decisions.

As expected, allowing for a variable bias in the threshold for switching from one choice to another did lead to higher levels of matching between the model and human subjects testing. Also, as one would expect, tailoring the value of the threshold to the task and to the human subject led to better predictive capabilities in the model than using a single global value for all tasks and subjects. Further, when allowing for a time-varying bias during the experiment, a clear pattern of discrete changes in value was observed suggesting that particular decision-making policies were used for a period of time, then altered.

These results suggest that the deterministic model studied here does have a nontrivial relation to the decision-making strategies used by human subjects in these simple tasks, particularly when the model is fit to the particular person. As discussed in the introduction to the work, actual decision-making policies are almost certainly a mix of deterministic and stochastic elements. Integrating appropriate stochastic terms with the model here is the subject of ongoing work. In particular, we are utilizing the deterministic model here as the basis of a finite state machine with probabilistic transitions between the modes in an MDP framework. The improvements from the model extensions indicate that the model can potentially be adjusted to provide a good fit with human decision making in TAFC tasks. Ongoing work is directed at these improvements. Efforts are also being pursued to produce real-time identification of the value of δ for each subject and to investigate the correlation between this type of parameter with aspects of human psychology such as risk attitude and the effect of deadline pressure into the model. Real-time tracking of such characteristics would potentially provide a means of improving human interaction with autonomous systems by incorporating the parameter as a feedback term in the interaction of the autonomous portion of the system with the human operator. ■

REFERENCES

[1] D. Baronov and J. Baillieul, "Search decisions for teams of automata," in *Proc. 47th IEEE Conf. Decision Control*, 2008, pp. 1133–1138.

[2] M. Cao, A. R. 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*, 2008, pp. 1127–1132.

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

[4] R. Parasuraman, M. Barnes, and K. Cosenzo, "Adaptive automation for human-robot teaming in future command and control systems," *Int. C2 J.*, vol. 1, no. 2, pp. 43–68, 2007.

[5] K. Stubbs, D. Wettergreen, and P. J. Hinds, "Autonomy and common ground in human-robot interaction: A field study," *IEEE Intell. Syst.*, vol. 22, no. 2, pp. 42–50, Mar./Apr. 2007.

[6] R. Simmons, S. Singh, F. Heger, L. M. Hiatt, S. C. Koterba, N. Melchior, and B. P. Sellner, "Human-robot teams for large-scale assembly," in *Proc. NASA Sci. Technol. Conf.*, 2007.

[7] N. Schurr, J. Marecki, M. Tambe, P. Scerri, N. Kasinadhuni, and J. Lewis, "The future of disaster response: Humans working with multiagent teams using DEFACTO," in *Proc. AAAI Spring Symp. AI Technol. Homeland Security*, 2005.

[8] 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.

[9] R. J. Herrnstein, "Rational choice theory: Necessary but not sufficient," *Amer. Psychol.*, vol. 45, no. 3, pp. 356–367, 1990.

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

[11] 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, no. 4, pp. 700–765, 2006.

[12] 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.

[13] J. R. Busemeyer and J. T. Townsend, "Fundamental derivations from decision field theory," *Math. Social Sci.*, vol. 23, no. 3, pp. 255–282, 1992.

[14] A. Diederich, "Dynamic stochastic models for decision making under time constraints," *J. Math. Psychol.*, vol. 41, no. 3, pp. 260–274, 1997.

[15] P. L. Smith, "Stochastic dynamic models of response time and accuracy: A foundational primer," *J. Math. Psychol.*, vol. 44, no. 3, pp. 408–463, 2000.

[16] A. Diederich and J. R. Busemeyer, "Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time," *J. Math. Psychol.*, vol. 47, no. 3, pp. 304–322, 2003.

[17] F. A. Sonnenberg and J. R. Beck, "Markov models in medical decision making," *Med. Decision Making*, vol. 13, no. 4, p. 322, 1993.

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

[19] M. Cao, A. Stewart, and N. E. Leonard, "Convergence in human decision-making dynamics," *Syst. Control Lett.*, vol. 59, no. 2, pp. 87–97, 2010.

[20] A. Stewart, M. Cao, A. Nedic, D. Tomlin, and N. E. Leonard, "Towards human-robot teams: Model-based analysis of human decision making in two-alternative choice tasks with social feedback," *Proc. IEEE*, vol. 100, no. 3, pp. 751–775, Mar. 2012, DOI: 10.1109/JPROC.2011.2173815.

[21] C. Woodruff, K. A. Morgansen, L. Vu, and D. Tomlin, "Modeling and evaluation of decision-making dynamics in sequential two-alternative forced choice tasks," in *Proc. 49th IEEE Conf. Decision Control*, 2010, pp. 3802–3807.

[22] P. Squire, J. G. Trafton, and R. Parasuraman, "Human control of multiple unmanned vehicles: Effects of interface type on execution and task switching times," in *Proc. 1st Conf. Human-Robot Interaction*, 2006, pp. 26–32.

[23] B. Trouvain and H. L. Wolf, "Evaluation of multi-robot control and monitoring performance," in *Proc. IEEE Int. Workshop Robot Human Interactive Commun.*, Berlin, Germany, Sep. 2002, pp. 111–116.

[24] S. R. Dixon, C. D. Wickens, and D. Chang, "Unmanned aerial vehicle flight control: False alarms versus misses," in *Proc. Human Factors Ergonomics Soc. Annu. Mtg.*, New Orleans, LA, 2004, pp. 152–156.

[25] H. A. Ruff, G. L. Calhoun, M. H. Draper, J. V. Fontejon, and B. J. Guilfoos, "Exploring automation issues in supervisory control of multiple uavs," in *Proc. Human Perf. Situation Awareness Autom. Conf.*, Daytona Beach, FL, 2004, pp. 218–222.

[26] C. E. Nehme, "Modeling human supervisory control in heterogenous unmanned vehicle systems," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, Feb. 2009.

[27] 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*, 2008, pp. 1115–1120.

[28] H. H. Kelley, J. W. Thibaut, R. Radloff, and D. Mundy, "The development of cooperation in the 'minimal social situation'," *Psychol. Monographs*, vol. 76, no. 19, pp. 1–19, 1962.

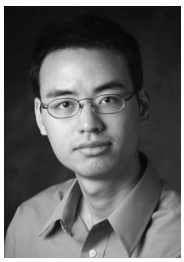
[29] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[30] J. D. Cohen, S. M. McClure, and A. J. Yu, "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 362, no. 1481, pp. 933–942, 2007.

ABOUT THE AUTHORS

Caleb Woodruff received the B.S.E. degree in mechanical engineering from Walla Walla University, College Place, WA, in 2009 and the M.S.A.A. degree in aeronautical and astronautical engineering with specialization in control theory from the University of Washington, Seattle, in 2011.

He has worked in the field of small unmanned aerial vehicles (UAVs) in summers since high school at Hood Technology Inc., Hood River, OR, where he is currently employed in a yearlong internship. His current interests include sensor path planning, manned imaging platforms, and hobby built UAVs.



Linh Vu (Member, IEEE) received the B.Eng. degree (with honor class I) in electrical engineering from the University of New South Wales, Sydney, N.S.W., Australia, in 2002 and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2003 and 2007, respectively.

From 2002 to 2007, he was a Research Assistant in the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign under the guidance of Prof. D. Liberzon. From 2007 to 2010, he was a Researcher with the Nonlinear Dynamics and Control Lab, Department of Aeronautics and Astronautics, University of Washington, working with Prof. K. A. Morgansen. He is currently with the Department of International and Postgraduate Study, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. His research interests include analysis and synthesis of switched systems, adaptive control, nonlinear control, and distributed multiagent systems.

Dr. Vu (together with Prof. Morgansen) is the recipient of the 2010 O. Hugo Shuck best paper award from the American Automatic Control Council (AACC).

Kristi A. Morgansen (Senior Member, IEEE) received the B.S. (*summa cum laude*) and M.S. degrees in mechanical engineering from Boston University, Boston, MA, in 1993 and 1994, respectively, and the S.M. degree in applied mathematics and the Ph.D. degree in engineering sciences from Harvard University, Cambridge, MA, in 1996 and 1999, respectively.

After receiving the Ph.D. degree, she was first a Postdoctoral Scholar and then a Senior Research Fellow in Control and Dynamical Systems and Mechanical Engineering at the California Institute of Technology, Pasadena. In August 2002, she joined the faculty of the Department of Aeronautics and Astronautics, University of Washington, Seattle, where she is currently an Associate Professor with tenure. From 2002 to 2007, she held the chaired position of Clare Boothe Luce Assistant Professor of Engineering at the University of Washington. Her research interests include nonlinear and coordinated control systems, bioinspired sensing and actuation, fin-based propulsive methods, control of coordinated systems with communication constraints, vision-based sensing for state estimation, and development of integrated human and autonomous multivehicle systems.

Prof. Morgansen received a National Science Foundation (NSF) CAREER Award in 2003 and the 2010 O. Hugo Schuck Award for Best Paper in the Theory Category in the 2009 American Control Conference.

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

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

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