Strategic Advice Provision in Repeated Human-Agent Interactions

Amos Azaria[1] and **Zinovi Rabinovich**[1] and **Sarit Kraus**[1,2] and **Claudia V. Goldman**[3] and **Ya'akov Gal**[4]
[1] Department of Computer Science, Bar-Ilan University, Ramat Gan 52900, Israel
[2] Institute for Advanced Computer Studies University of Maryland, MD 20742
[3] General Motors Advanced Technical Center, Herzliya 46725, Israel
[4] Department of Information Systems Engineering, Ben-Gurion University of the Negev, Israel
{azariaa1,sarit}@cs.biu.ac.il, zr@zinovi.net, claudia.goldman@gm.com, kobig@bgu.ac.il

Abstract

This paper addresses the problem of automated advice provision in settings that involve repeated interactions between people and computer agents. This problem arises in many real world applications such as route selection systems and office assistants. To succeed in such settings agents must reason about how their actions in the present influence people's future actions. This work models such settings as a family of repeated bilateral games of incomplete information called "choice selection processes", in which players may share certain goals, but are essentially self-interested. The paper describes several possible models of human behavior that were inspired by behavioral economic theories of people's play in repeated interactions. These models were incorporated into several agent designs to repeatedly generate offers to people playing the game. These agents were evaluated in extensive empirical investigations including hundreds of subjects that interacted with computers in different choice selections processes. The results revealed that an agent that combined a hyperbolic discounting model of human behavior with a social utility function was able to outperform alternative agent designs, including an agent that approximated the optimal strategy using continuous MDPs and an agent using epsilon-greedy strategies to describe people's behavior. We show that this approach was able to generalize to new people as well as choice selection processes that were not used for training. Our results demonstrate that combining computational approaches with behavioral economics models of people in repeated interactions facilitates the design of advice provision strategies for a large class of real-world settings.

Introduction

As computers become ubiquitous, settings in which they make decisions with people over time are becoming increasingly prevalent. Many of these settings require computer agents to generate advice to their human users about which decisions to take in a way that guides their behavior. Such settings arise in a variety of application domains such as hospital care-delivery systems, negotiation training or route-navigation systems. Although computers and people in these domains share some goals, such as completing the user's tasks, their goals may also conflict. For example, consider an environmentally-conscious route-selection system that advises drivers about commuting routes daily. The system possesses information about traffic jams and road conditions that is not available to the driver who makes the decision which route to take. Both system and driver wish to reach the destination safely. However, the driver may prefer quicker routes, while the system cares about reducing the driver's impact on the environment. Another example involves a decision-support system for doctors for the purpose of recommending medical treatments to patients. The system may have knowledge of a new highly effective antibiotic, but will suggest a more traditional treatment for the patient in order to alleviate drug resistance in the population.

This paper addresses problems central to the design of advice provision strategies for computer agents that interact with people in repeated settings. We model these interactions as a family of repeated games of incomplete information called *choice selection processes* comprising a human and computer player. Both of the participants in a choice selection process are self-interested. The computer possesses private information regarding the states of the world which influences both participants' outcome; this information is not known to the person. At each round, the computer suggests one of several choices to the person, and the person then selects his or her choice, which may or may not correspond to the computer's suggestion. This choice affects the outcome for both the person and the computer agent.

For an agent to be successful in such interactions, it needs to generate advice that is likely to be accepted by people, while still fulfilling the agent's goals. The design of such advice provision strategies is computationally challenging for several reasons. First, the agent needs to reason about the potential effect of the proposed action on its future interactions with people. For instance, suggesting routes that are significantly more beneficial to the agent may cause the person to ignore its future recommendations. Second, people's decision-making deviates from rational choice theory and is affected by a variety of social and psychological factors (Camerer 2003). For instance, some people may prefer certain routes due to their past experience and may be reluctant to adopt new, possibly preferable alternatives. Lastly, people have been shown to discount the advice they receive from experts (Bonaccio and Dalal 2006).

To address these challenges, we designed several models of human behavior in choice selection processes that incorporated quantal response, exponential smoothing, and hyperbolic discounting theories from behavioral economics. The parameters of these models were estimated using maximum likelihood, and their predictive power was measured on a sampled set of human play using ten-fold cross-validation. The best model—a combination of hyperbolic discounting and quantal response—was incorporated into an agent that was evaluated in an extensive study involving hundreds of people. The study consisted of a repeated setting that is analogous to choice selection processes, in which a person is asked to choose a route to work from a set of possible candidates. The travel time and the fuel consumption of each road varies due to traffic, and is known to the computer (but not the person). At each round, the computer suggests one of the routes to the person. The person's goal is to minimize travel time while the agent's goal is to minimize fuel consumption.

To propose routes, the agent used a social utility approach which considered the costs for both agent and person when making suggestions. This agent was compared to several alternative strategies for advice generation, including an agent that approximated the optimal strategy using a continuous Markov Decision Process. Each round in the evaluation varied the road and fuel conditions using models from the transportation engineering literature of real-world road conditions. The agents were evaluated in simulations using thousands of rounds as well as in studies comprising actual people. Results show that the social utility agent was consistently able to outperform all other agent strategies using two different models of human behavior. This work is the first to design a computer agent for generating advice to people in repeated settings, and demonstrates the efficacy of using behavioral economic models when generating advice.

Related Work

Past work on advice provision spans the computational and social sciences. Game theory researchers have studied persuasion games (Milgrom and Roberts 1986; Crawford and Sobel 1982), in which a Sender player needs to decide how much information to disclose to a Receiver player to influence the Receiver's strategy in a way that benefits the Sender. The majority of works on persuasion games study one-shot interactions (Sher 2011; Glazer and Rubinstein 2006). A notable exception is the work by Renault et al. (2011) who considered repeated interactions that follow a Markov chain observed solely by the Sender. The Receiver cannot observe its utility until the end of the multi-period interactions. All of these works make the strong assumption that people follow equilibrium strategies. However, agents using equilibrium approaches to interact with people are not successful in repeated settings (Hoz-Weiss et al. 2008; Peled, Gal, and Kraus 2011).

Models for predicting users' ratings have been proposed that are used by recommendation systems to advise their users (See Ricci et al. (2011) for a recent review). Most works in this realm have only considered the utility of the system and have not modeled the user's reactions to its actions over time. One exception is the work by Shani et al. (2005), which uses a discrete-state MDP model to maximize the system utility function taking into account the future interactions with their users. This approach is not applicable to our domain, which incorporates a continuous state space and histories of arbitrary length. In addition, the model they use does not consider the possible effects of the recommendations on user's future behavior. Viappiani and Boutilier (2009) propose an approach to recommender systems that incorporates explicit utility models into the recommendation process. They assume that the utility for the user and the system are the same, while in our work the systems utility may be unrelated to that of the user.

Another method which diverts from standard recommender system is the work presented by (Sarne et al. 2011). They attempt to facilitate people's decision making process by modifying the presentation of the problem. However, once again they assume the utility of the user and the system are identical.

In previous work (Azaria et al. 2011) we allowed the system to reveal partial information in order to encourage the user to take a certain action, rather than explicitly recommending an action (as in current work).

Lastly, we mention recent work in AI (Azaria et al. 2012) that proposed an advice provision strategy for a one-shot interaction in which both participants had complete information about the state of the world. Our work extends this approach to the more challenging, yet realistic setting of repeated interactions under incomplete information.

Choice Selection Processes

A choice selection process is a repeated game of incomplete information played by two players, a Receiver and a Sender. The Receiver chooses an action a out of a possible set A. To illustrate, in the route-selection example the set of actions comprises the possible roads that the Receiver can choose. The state of the world $v = (v_1, \ldots, v_{|A|})$ is a multivariate continuous random variable. Each element v_i of v represents information about the world that affects players' outcomes given that the Receiver chooses action a_i. For example, an element v_i can incorporate the expected time and fuel consumption for using road i and v incorporates the time and fuel consumption for all possible roads. At each round t, the Sender observes the state of the world v and can suggest to the Receiver to take one of the actions in A before the Receiver makes its selection. The Receiver does not observe v nor does it know its underlying distribution. The cost $c_s(v, a)$ to the Sender and the cost $c_r(v, a)$ to the Receiver correspond to the fuel consumption and time in our example, are determined by the state of the world and the Receiver's action. Their interactions continue repeating with a constant probability γ (i.e., the discount factor).

Modeling Human Receivers

Because the human does not know the distribution over the states of the world, its decision problem can be analogously described as a Multi Armed Bandit Problem (MAB) (Auer et al. 1995), in which there are $|A| + 1$ arms (one for each action, and one for following the advice of the Sender). We present several candidate models for describing human Receiver behavior that combine heuristics from the MAB literature with theories from the behavioral economics literature.

We begin with the following notation. The behavior in a selection process at a round t is represented by a tuple $h^t = (a^t, c^t, d^t)$ where a^t is the Receiver's action at round t, $c^t = (c^t_r, c^t_s)$ is the cost to the Receiver and Sender at t, and d^t is the advice provided by the Sender at t (prior to the Receiver choosing a^t). Since the human player does not know the actual state of the world, and in particular the costs of all actions in each round, we use the notion of subjective cost to model its reasoning when considering which action to take. Given behavior h^t at time t, we define the subjective cost to the Receiver for taking action a at time t, denoted $sc_r(a, t \mid h^t)$, to equal the cost c^t_r when $a = a^t$ (i.e., the Receiver chose action a^t at time t); otherwise it equals some default value K, as the human does not know what cost would have been incurred by taking action a for rounds that it was not chosen. For example, suppose that the Receiver chose to use route 66 on day 1 and incurred a 45 minute commute. The subjective cost to the Receiver for using route 66 on day 1 equals 45 minutes, while the subjective cost for using any other route equals the default value. Similarly, we define the subjective cost to the Receiver for taking advice d at time t, denoted $sc_r(d, t \mid h^t)$, to equal the cost c^t_r when $a^t = d$ (i.e., the Receiver followed the Sender's advice), or a default value.

We now generalize the notion of subjective cost to include behavior over multiple rounds. Let $z(a, t \mid h^{1,t-1})$ denote the aggregate subjective cost to the Receiver at rounds 1 through $t-1$ for taking action a. The models we describe below differ in how they aggregate the Receiver's subjective costs over time. We begin with two models in which receivers discount their past costs higher than their present costs. In the *hyperbolic discounting model* (Chabris, Laibson, and Schuldt 2006; Deaton and Paxson 1993), the discount factor δ falls very rapidly for short delay periods, but falls slowly for longer delay periods. For example, consider a driver that took a new route to work on Monday which happened to take an hour longer than the route on Friday. According to hyperbolic theory, the relative difference between the commute times will be perceived to be largest during the first few days following Monday. However, as time goes by, the perceived difference between the commute times will diminish.

$$z(a, t \mid h^{1,t-1}) = \sum_{t'<t} \frac{sc_r(a \mid h^{t'})}{\delta \cdot (t - t')} \tag{1}$$

In the *Exponential Smoothing* model (Gans, Knox, and Croson 2007), the discount factor δ is constant over time, meaning the perceived difference between the commute times will stay the same over time. The subjective cost for the Receiver is defined as follows. If $a^{t-1} = a$ (the Receiver took action a at time $t-1$) then we have

$$\begin{aligned} z(a, t \mid h^{1,t-1}) =& \delta \cdot sc_r(a \mid h^{t-1}) + \\ & (1-\delta) \cdot z(a, t-1 \mid h^{1,t-2}) \end{aligned} \tag{2}$$

If $a^{t-1} \neq a$, the Receiver does not update its aggregate subjective cost for action a, and we have

$$z(a, t \mid h^{1,t-1}) = z(a, t-1 \mid h^{1,t-2}) \tag{3}$$

If $t = 1$ then $z(a, t \mid h^{1,t-1})$ equals a default value L for any a.

In the *Short Term Memory* model, the Receiver's valuation is limited to 7 past rounds, (the "magic number" commonly associated with human short term memory (Miller 1956; Lisman and Idiart 1995)). The subjective cost for the Receiver is defined as follows:

$$z(a, t \mid h^{1,t-1}) = \sum_{t-7 \leq t' < t} sc_r(a, t' \mid h^{1,t'-1}) \cdot \frac{1}{7} \tag{4}$$

If $t < 7$, then the summation only spans rounds $1, \ldots, t$, and the denominator is replaced by t.

Lastly, we also considered a baseline *Soft Max* model in which the aggregate subjective cost of the Receiver for any action is the average cost of taking this action in past rounds, with no discount factor.

To model the Receiver's action at time t, we adopted the quantal response theory from behavioral economics (Haile, Hortasu, and Kosenok 2008) that assigns a probability of choosing an action a that is inversely proportional to the aggregate subjective cost of that action given the history (i.e. $z(a, t \mid h^{t-1})$). The Receiver is modeled to prefer actions associated with lower subjective costs. However, with some probability, the Receiver may still choose actions that are more costly. The probability of the action a also depends on the term $z(d^t, t \mid h^{t-1})$, which is the aggregate subjective cost to the Receiver from following the advice of the Sender at rounds $1, \ldots, t-1$. Formally, the probability distribution that the Receiver takes action a^t at round t given behavior at past rounds $h^{1,t-1}$ is

$$\begin{aligned} & P(a, t \mid h^{1,t-1}, d^t) = \\ & \frac{e^{-\lambda \cdot z(a^t, t \mid h^{1,t-1})} + S}{e^{-\lambda \cdot z(d^t, t \mid h^{1,t-1})} + \sum_{a \in A} e^{-\lambda \cdot z(a, t \mid h^{1,t-1})}} \end{aligned} \tag{5}$$

Where S is set to equal $e^{-\lambda \cdot z(d^t, t \mid h^{1,t-1})}$ when $a = d^t$, and zero otherwise; λ is a smoothing parameter.

Generating Strategies for the Sender

In this section we formally define the problem of finding the optimal strategy for the Sender, and present several approximate solutions to the problem. Both the formal optimal policy and its approximations treat the model of the Receiver's behavior as a parameter.

To formally define an optimal strategy of a Sender, we first represent the Sender's decision making process as a continuous MDP. To represent the selection process from the Sender's point of view as an MDP, we define the set of world states for the MDP as follows.[1] For any time t, there is a corresponding world state for any state $v^t \in V$ and history sequence $h^{1,t} \in H^{1,t}$. The set of all such world states is denoted as $S^t = \{(v^t, h^{1,t-1}, t) \mid v^t \in$

[1] We use the term "world state" to disambiguate the states of an MDP from those of a selection process.

$V, h^{1,t} \in H^{1,t}\}$. The set of possible world states is defined as $S = \cup_{t=1,\dots,\infty} S^t \cup \{s_a^{term} \mid a \in A\}$. The termination states s_a^{term} represent the end of a selection process after the Receiver executed action a, and will allow a coherent definition of the MDP cost. The set of actions for the Sender is the set $|A|$ of actions in the selection process. Let $s^t = (v^t, h^{1,t-1}, t)$ and $s^{t+1} = (v^{t+1}, h^{1,t}, t+1)$, where $h^{1,t} = h^{1,t-1} \circ \langle a^t, (c_r(v^t, a^t), c_s(v^t, a^t), d^t \rangle$, be two world states. Then the cost function for the MDP, denoted c^{MDP}, is defined as $c^{MDP}(s_a^{term} \mid s^t, d^t) = c^{MDP}(s^{t+1} \mid s^t, d^t) = c_s(v^t, a^t)$ and zero in all other cases. Notice that the cost depends on the Receiver's action that is encoded in the target state, specifically h^t in s^{t+1} and the action index in case of s_a^{term}.

The transition function of the MDP describes the progress of the selection process. In the case that the selection process terminates with probability $1 - \gamma$, we set $P(s_a^{term} \mid s^t, d^t) = (1-\gamma) P(a^t \mid h^{1,t-1}, d^t, t)$, where s^t as above and $P(a^t \mid h^{1,t-1}, d^t, t)$ is the probability that the Receiver chooses action a^t at time t given the history $h^{1,t-1}$ and that the Sender offers d^t. In the case that the selection process continues with probability γ, we set $P(s^{t+1} \mid s^t, d^t) = \gamma \cdot P(a^t \mid h^{1,t-1}, d^t, t) \cdot P(v^{t+1})$, where s^{t+1} as above and $P(v^{t+1})$ is the probability that the selection process state v^{t+1} will occur. To complete the transition function we also set $P(s_a^{term} \mid s_{a'}^{term}, d^t) = 1$ if $a = a'$ and zero otherwise. Finally, the initial state of the MDP is sampled from the world states subset $\{(v, \emptyset, 1) \mid v \in V\}$ according to $P(v)$, and the optimality criterion is set to be the minimization of the expected accumulated cost.

Solving the continuous MDP described above yields an optimal policy for the Sender given a model of the Receiver, $P(a^t \mid h^{1,t-1}, d^t, t)$. However, the world states of the MDP incorporate the continuous state of the selection process and discrete histories of arbitrary length, which makes the MDP structure too complex to be solved exactly. In addition, we cannot use existing approximation algorithm, which assume a finite state space (Marecki, Koenig, and Tambe 2007), partition the state space (Feng et al. 2004), or use kernel-based methods (Ormoneit and Sen 2002), due to the mixture of the continuous component (selection process state) and an arbitrarily large discrete component (action and advice history) of the world state.

We therefore devise two approximate solutions by a-priory limiting the space of possible strategies available to the Sender and then, *within* the limited sub-set of strategies, find the optimal one. The first approximation approach allows only for an optimal strategy with respect to a single-step policy iteration step. Specifically, we calculate

$$d^{*,t} = \arg\min_{d^t \in A} \int\limits_{s^{t+1} \in S} \left(c^{MDP}(s^{t+1} | s^t, d^t) + V(s^{t+1}) \right) \cdot P(s^{t+1} | s^t, d^t) ds^{t+1} \quad (6)$$

where $V(s^{t+1})$ is the expected accumulated cost of the random strategy, i.e. where the Sender offers any $d \in A$ with probability $1/|A|$. We estimated $V(\cdot)$ using Markov Chain Monte Carlo sampling in a manner similar to that of (Lanctot et al. 2009).

The second approximation approach allows the Sender to generate advice by minimizing the weighted normalized cost to both the Sender and the Receiver at each round. This approach is inspired by social preferences models that have been shown to be effective for repeated human-computer decision-making (Gal et al. 2012). Given a state v, the offer chosen by the Sender is defined as follows

$$d^{*,t} = \arg\min_{d^t \in A} (1-w) \cdot \frac{1}{N_r} \cdot (c_r(d^t, v^t)) + w \cdot \frac{1}{N_S} \cdot (c_s(d^t, v^t)) \quad (7)$$

where w is a constant weight, and N_r and N_S are normalizing factors for the costs to the Receiver and Sender. For a given world state v and history h^t, we can define the Sender's expected cost $EC_s(v, h^t)$ for action $d^{*,t}$ as

$$EC_S(v, h^t) = \sum_{a \in A} P(a^t \mid h^{1,t-1}, d^{*,t}, t) c_s(a, v) \quad (8)$$

The weight w is chosen to minimize the Sender's costs when summing over all world states and histories, as defined below.

$$\arg\min_{w' \in [0,1]} \int_v \sum_{t=1}^{\infty} \sum_{h^t} \gamma \cdot P(v) \cdot EC_S(v, h^t) dv \quad (9)$$

The social weight approach explicitly reasons about the trade-offs between the costs to both participants in the selection process.

Empirical Methodology

Our empirical methodology comprises a family of selection processes that are analogous to a route-selection task between a driver (a human Receiver) and a navigation system (an agent Sender). At each round of the interaction, the driver needs to choose one of 4 possible routes to get to work. The system can advise the driver to take one of the routes before the driver makes a choice. The road conditions (i.e., travel time and fuel consumption) constitute the state of the world, and vary due to traffic and maintenance. They are unknown to the driver when it makes its decision. The driver's goal is to minimize the travel time over all rounds, and the system's goal is to reduce fuel consumption over all rounds. After the driver chooses a route, both participants incur a cost which depends on the road conditions for the chosen route. At this point the interaction continues to the next round with probability 0.96. The conditions for the roads at each round are sampled from a joint distribution that is known to the agent, but not to the driver.

We modeled the fuel consumption and travel time using a multivariate log-normal distribution. We used two different settings in the study. In the first, we generated road conditions using a model from the transportation engineering literature (Ahn et al. 2002) which formalized a dependency between time and fuel consumption as a function of the expected car accelerations in the road. In the second condition, the time and fuel consumption were independent from each other. Table 1 shows the mean fuel consumptions (in liters)

Table 1: Road Condition Statistics for Independent and Dependent Settings

environment	mean fuel consumption (liters)	mean travel time (minutes)
ind.	4.06	70
dep.	7.05	80.24

and commuting time (in minutes) for all 9375 rounds used in the study. As shown by the table, the road conditions generated for the dependent setting were considerably more costly than those generated for the independent setting.

We enlisted 375 subjects using Amazon Mechanical Turk, all of which were provided with a description of the route selection task. Subjects were told that the probability of a new round was 0.96. However, to standardize comparison of results, each subject actually played 25 rounds. This number is the expected number of rounds given that the discount factor was 0.96, and also, the average number of commuting days in one month. The actual number of rounds was not revealed to subjects (nor the computer agents).[2] Subjects were paid a bonus proportional to the average travel time (the lower the travel time the higher the bonus). All subjects were provided with an explanation of the game and its details, which we described in the beginning of this section.

Model Selection and Agent Design

To choose between the various models of the Receiver, we collected 2250 rounds from 90 subjects to train and evaluate the short-term memory (ST), hyperbolic discounting (Hyper), SoftMax (SM), and Exponential Smoothing (ES) models that were described earlier. For each of these models, we estimated the maximum-likelihood (ML) value of the parameters using sampling, and computed the fit-to-data of the test set using the ML values. All results reported throughout the section were confirmed to be statistically significant using single factored analysis of variance tests (ANOVA) for $p < 0.05$. Table 2 presents the fitness of the models employing ten-fold-cross-validation (lower values indicate a better fit of the model). As shown by the table, the Hyper model, which modeled the Receiver using hyperbolic discounting theory (Equations 1 and 5) exhibited a higher fit-for-data than all other models of human Receivers.[3]

We hypothesized that the use of the social utility approach will lead to best performance for the agent Sender, measured in terms of fuel consumption. To evaluate this hypothesis, we used different agent designs for generating offers to people which incorporate the decision-making strategies that were described in the previous section. Specifically, we used an agent that incorporated the social utility approach to make offers, termed Social agent for Advice Provision (*SAP*), and an agent using the MDP model to make offers, termed *MDP*. We also employed two baseline agents, *Random* which offered roads with uniform probability and *Silent* which didn't provide any advice.

Table 2: Fit-to-data of Different Receiver Models (lower is better)

model	d.f.	Log-Like.
SoftMax	1	178.5
ES	2	172.2
hyper	2	**169.4**
short memory	1	186.9

Table 3: Simulation results comparing agent strategies

human model	agent strategy	fuel (liters)	time (minutes)
hyper	Random	6.120	64.40
	Silent	6.297	63.04
	MDP	5.792	65.92
	SAP	**5.520**	64.54
$\epsilon-$greedy	Random	7.046	58.08
	Silent	7.104	57.68
	MDP	6.812	59.26
	SAP	**6.432**	55.84

We evaluated these agent designs in simulation as well as in experiments involving new people. The simulation studies consisted of sampling 10,000 road instances according to the distribution over the fuel consumption and travel time for independent settings. As an alternative to the hyperbolic discounting model, we also considered an approach using an $\epsilon-$greedy strategy to describe Receiver behavior. This strategy is commonly used to solve Multi Armed Bandit problems (Vermorel and Mohri 2005), which describes the choice selection problem from the point of view of the Receiver. This strategy provides a rational baseline that seeks to minimize travel time for Receivers over time. Table 3 presents results of the simulation. We compared the fuel consumption costs incurred by the different Sender agents for each model used to describe human behavior. As shown in Table 3, the cost accumulated by the SAP agent using the hyperbolic discounting model was 5.52 liters (shown in bold), which was significantly lower than the cost to all other agents using the hyper models to describe human behavior. Similarly, the cost accumulated by the SAP agent using the $\epsilon-$greedy model were significantly lower than the cost to all other agents using $\epsilon-$greedy models.

Evaluation with People

Given the demonstrated efficacy of the SAP agent in the simulation described above, we set out to evaluate the ability of the SAP agent to generalize to new types of settings and new people. We hypothesized that a SAP agent using the hyperbolic discounting model to describe Receiver behavior would be able to improve its performance when compared to the SAP agent using the $\epsilon-$greedy model. We randomly divided subjects into one of several treatment groups. Sub-

[2] All of the study procedures were authorized by the ethics review board of the corresponding institutions.

[3] For all models, we set the default value K to equal the mean travel time of the road associated with the highest commuting time, representing an upper bound for the cost to the Receiver.

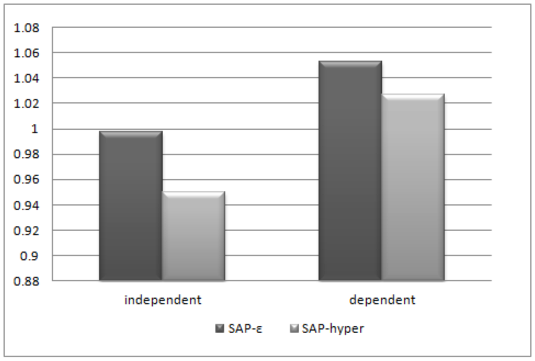
Table 4: Performance Results against People

method	selfishness	fuel	time	acceptance
Silent	–	6.20	64	–
Receiver	0	6.44	56.6	63.6%
Sender	1	5.88	64.32	31.0%
SAP-ϵ	0.29	5.76	56.6	70.8%
SAP-hyper	0.58	**5.08**	64.8	52.6%

Table 5: Generalizing to new people and new road conditions

time & fuel settings	agent-strategy	selfishness	fuel	time	acceptance
ind.	SAP-ϵ	0.33	4.048	49.98	67.0%
ind.	SAP-hyper	0.66	**3.858**	55.48	41.9%
dep.	SAP-ϵ	0.26	7.423	70.96	63.0%
dep.	SAP-hyper	0.68	**7.236**	73.58	37.3%

Figure 1: Normalized performance results for SAP agents



jects in the *Silent* group received no advice at all. Subjects in the *SAP-hyper* group were advised by the SAP agent using a hyperbolic model to describe Receiver behavior. Subjects in the *SAP-ϵ* group were advised by the SAP agent that used an $\epsilon-$greedy strategy to describe Receiver behavior. Subjects in the *Receiver* group were consistently advised to choose the road that was most beneficial to them, (i.e., associated with the lowest travel time). Lastly, subjects in the *Sender* group were consistently advised to choose the road which was best for the Sender (i.e., associated with the lowest fuel consumption).

Table 4 shows results for evaluating the models with new people on the same instances used to train the models. All of these instances included independent road conditions. The performance for agents and for people is measured in terms of overall fuel consumption and commuting time, respectively. The "selfishness" column measures the degree to which the agent was self-interested (the weight w in Equation 7). As shown by table 4, the SAP-hyper agent significantly outperformed all other agent-designs, accumulating a cost of 5.08 liters (shown in bold). The best performance for people (travel time of 56.6 minutes) was obtained when using an agent that only considered people's costs (*Receiver*) as well as the $\epsilon-$greedy agent. The acceptance rates for the SAP-hyper were lower than those for SAP-ϵ, which we attribute to the higher degree of selfishness for the SAP-hyper agent. Surprisingly, the acceptance rate for SAP-ϵ was higher than that of the Receiver agent, whose degree of selfishness was 0, and consistently recommended the route that was best for people. We hypothesize that this may be caused by an unintended "too-good-to-be-true" signaling effect that is perceived by people.

Table 5 reports results when comparing between SAP-ϵ and SAP-hyper agents on new games and new people when sampling road conditions from both independent and dependent distributions. As shown by the table, the SAP-hyper agent was also able to outperform the SAP-ϵ agent when generalizing to new settings, although the differences in performance were smaller. Figure 1 normalizes the performance of both SAP agents across the independent and dependent settings by dividing the costs incurred in the independent and dependent settings by the average costs over all possible actions. As shown by the Figure, the costs incurred by both SAP agents on the dependent settings were higher than the independent setting, which we attribute to the realistic and more challenging characteristics of these instances.

One may be bothered by the relatively low user acceptance rate or by the relatively poor user performance for SAP-hyper. This may raise a concern that SAP might not perform as well when longer interactions are expected. Recall that the agent's goal was only to minimize its own cost. Although the agent did consider the user's cost and thus its satisfaction, it was considered a means to an end in order to minimize the agent's overall cost. If the system expects a longer period of interaction with the user (i.e. greater γ), the user's satisfaction will be more important to the agent, and therefore the social weight will be balanced towards the users benefit (causing an increase in user acceptance rate and performance). Furthermore, if user satisfaction is important to the agent on its own, it can be explicitly added to the agent's utility. However, we chose a more confrontational setting to demonstrate the methods efficacy.

We conclude this section with two illustrative examples of the reasoning in use by the SAP-hyper agent. In the first example, one of the roads incurs very low cost to the agent (3 liters), but has an extremely high cost for the person (43 minutes). In this example, the SAP-hyper agent recommended the road that was associated with the *highest* cost to the agent (4.19 liters), but a very low cost to the person (18 minutes). By combining the social utility function with hyperbolic model, the SAP-hyper agent reasoned that this action could decrease its expected future costs. The person accepted this advice and chose the recommended route. In the next round, the agent advised the person to take a road with relatively high cost to the person (31 minutes) and very low cost to the agent (1.6 liters). This offer was again accepted, confirming the agent's reasoning.

Conclusions and Future Work

In this paper we consider a two player game, in which an agent repeatedly supplies advice to a human user followed by an action taken by the user which influences both the agent's and the user's costs. We present the Social agent for Advice Provision (SAP) which models human behavior combining principles known from behavioral science with machine learning techniques. We test different approaches for many of the SAP's assumptions, and show that the SAP agent outperforms all other alternatives. A further advantage of using the SAP-hyper agent, in addition to its demonstrated performance, is that its strategy does not depend on the history of interaction with current Receiver to generate an offer for the Sender. This makes it possible to deploy them in situations that are common to many route selection applications, where there is no knowledge of the number of times that users have used the system in the past.

In future work we intend to study scenarios where the user has other agents providing advice which have different cost functions than our agent. We will allow future users to receive partial information from other sources (analogous to a personal GPS or radio traffic reports) and allow them to turn on or off the advice received from our agent.

Acknowledgments

We thank Ofra Amir and Shira Abuhatzera for their helpful comments. We acknowledge K. Ahn, H. Rakha, A. Trani, and M. Van Aerde for their help in estimating fuel consumption and for their software. This work is supported in part by the following grants: ERC #267523, MURI #W911NF-08-1-0144, ARO #W911NF0910206 and #W911NF1110344, and Marie Curie #268362.

References

Ahn, K.; Rakha, H.; Trani, A.; and Van Aerde, M. 2002. Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *Journal of Transportation Engineering* 128(2):182–190.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 1995. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proc. of FOCS*, 322–331.

Azaria, A.; Rabinovich, Z.; Kraus, S.; and Goldman, C. V. 2011. Strategic information disclosure to people with multiple alternatives. In *Proc. of AAAI*.

Azaria, A.; Rabinovich, Z.; Kraus, S.; Goldman, C. V.; and Tsimhoni, O. 2012. Giving advice to people in path selection problems. In *Proceedings of AAMAS 2012*.

Bonaccio, S., and Dalal, R. S. 2006. Advice taking and decision-making: An integrative literature review and implications for the organizational sciences. *Org. Behavior and Human Decision Processes* Vol. 101(2):127–151.

Camerer, C. F. 2003. *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton Unv. Press. chapter 2.

Chabris, C.; Laibson, D.; and Schuldt, J. 2006. Intertemporal choice. *The New Palgrave Dictionary of Economics*.

Crawford, V., and Sobel, J. 1982. Strategic information transmission. *Econometrica* 50:1431–1451.

Deaton, A., and Paxson, C. 1993. Intertemporal choice and inequality. Work. P. 4328, NBER.

Feng, Z.; Dearden, R.; Meuleau, N.; and Washington, R. 2004. Dynamic programming for structured continuous markov decision problems. In *Proc. of UAI*.

Gal, Y.; Kraus, S.; Gelfand, M.; Khashan, H.; and Salmon, E. 2012. Negotiating with people across cultures using an adaptive agent. *ACM TIST* 3(1).

Gans, N.; Knox, G.; and Croson, R. 2007. Simple models of discrete choice and their performance in bandit experiments. *M&SOM* 9(4):383–408.

Glazer, J., and Rubinstein, A. 2006. A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics* 1:395–410.

Haile, P. A.; Hortasu, A.; and Kosenok, G. 2008. On the empirical content of quantal response equilibrium. *American Economic Review* 98(1):180–200.

Hoz-Weiss, P.; Kraus, S.; Wilkenfeld, J.; Andersend, D. R.; and Pate, A. 2008. Resolving crises through automated bilateral negotiations. *AIJ* 172(1):1–18.

Lanctot, M.; Waugh, K.; Zinkevich, M.; and Bowling, M. 2009. Monte carlo sampling for regret minimization in extensive games. In *Proc. of NIPS*, volume 22, 1078–1086.

Lisman, J. E., and Idiart, M. A. P. 1995. Storage of 7 ± 2 Short-Term Memories in Oscillatory Subcycles. *Science* 267:1512–1515.

Marecki, J.; Koenig, S.; and Tambe, M. 2007. A fast analytical algorithm for solving markov decision processes with real-valued resources. In *Proc. of IJCAI*.

Milgrom, P., and Roberts, J. 1986. Relying on the information of interested parties. *Rand J. of Economics* 17:18–32.

Miller, G. A. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review* 63(2):81–97.

Ormoneit, D., and Sen, S. 2002. Kernel-based reinforcement learning. *Machine Learning* 49(2):161–178.

Peled, N.; Gal, Y.; and Kraus, S. 2011. A study of computational and human strategies in revelation games. In *Proc. of AAMAS*.

Renault, J.; Solan, E.; and Vieille, N. 2011. Dynamic sender-receiver games. Unpublished manuscript.

Ricci, F.; Rokach, L.; Shapira, B.; and Kantor, P., eds. 2011. *Recommender Systems Handbook*. Springer.

Sarne, D.; Elmalech, A.; Grosz, B. J.; and Geva, M. 2011. Less is more: restructuring decisions to improve agent search. In *Proc. of AAMAS*, 431–438.

Shani, G.; Heckerman, D.; and Brafman, R. I. 2005. An MDP-based recommender system. *J. Mach. Learn. Res.* 6:1265–1295.

Sher, I. 2011. Credibility and determinism in a game of persuasion. *Games and Economic Behavior* 71(2):409 – 419.

Vermorel, J., and Mohri, M. 2005. Multi-armed bandit algorithms and empirical evaluation. In *Proc. of European Conference on Machine Learning*, 437–448. Springer.

Viappiani, P., and Boutilier, C. 2009. Optimal set recommendations based on regret. In *Proc. of the 7th Workshop on IT for Web Personalization & Recommender Systems*.