Human Behavior Models for Virtual Agents in Repeated Decision Making under Uncertainty

Ming Yin*
Harvard University
mingyin@fas.harvard.edu

Yu-An Sun
PARC
YuAn.Sun@xerox.com

ABSTRACT

To design virtual agents that simulate humans in repeated decision making under uncertainty, we seek to quantitatively characterize the actual human behavior in these settings. We collect our data from 800 real human subjects through a large-scale randomized online experiment. We evaluate the performance of a wide range of computational models in fitting the data by both conducting a scalable search through the space of two-component models (i.e. inference + selection model) and investigating a few rules of thumb.

Our results suggest that across different decision-making environment, an average human decision maker can be best described by a two-component model, which is composed of an inference model that relies heavily on more recent information (i.e. displays recency bias) and a selection model which assumes cost-proportional errors and reluctance to change in subsequent trials (i.e. displays status-quo bias). Additionally, while a large portion of individuals behave like the average decision maker, how they differ from each other is greatly influenced by the environment. These results imply the possibility of constructing agents with a single type of model that is robust against the context, and provide insights into adjusting heterogeneity among multiple agents based on the context.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General – *Cognitive simulation*

Keywords

Repeated Decision Making; Human-like Virtual Agents; Behavior Model; Cognitive Biases; Online Experiment

1. INTRODUCTION

Autonomous virtual agents that exhibit human-like behaviors have been widely applied in various domains, including multi-agent based simulations, serious games, arts and digital entertainment, education and virtual training [15, 4, 14]. One challenging problem with virtual agent design is how these agents should be modeled to resemble human behavior when they face uncertainty in decision-making, and in particular, when they have to make such decisions repeatedly. While most existing paradigms assume agents are rational decision makers hence display optimal behavior [28, 9], there is a large literature in behavioral economics and psychology showing that real humans are irrational as they are subject to cognitive biases and limitations [25, 23].

Thus, searching for a more realistic design of virtual agents in the setting of repeated decision making under uncertainty, a key question that needs to be addressed first is how real human beings *actually* behave in these conditions. For instance, how does a horse-racing gambler place bets on horses while observing their fluctuating performance over time; and how does a house owner choose her preferred electricity tariff scheme among multiple alternatives, such as the flat rate scheme (unit price is the same for usage in different time intervals) and different Time of Use (ToU) schemes (unit price is different for usage in different time intervals), while she periodically reviews her electricity bills?

To answer this question, in this paper, we aim at finding a computational model which best characterizes human behavior in repeated decision making under uncertainty. In particular, we focus on the following scenario which we refer to as the *environment learning problem*: There are N random variables in an environment (e.g. the amount of time for each of the N horses to finish a race, the electricity usage in each of the N time intervals of a day), and each random variable follows a *fixed* distribution that is *unknown* to the decision maker (DM). In each trial, the DM is first asked to choose among M options. Then, the DM observes an *independent* sample of *each* random variable for the trial. The DM's utility in a trial depends both on her choice and on the realized samples of all random variables in that trial. While the DM learns about the environment (i.e. random variable distributions) through the noisy samples, her objective is to maximize her cumulative utility in the whole decision period of T trials[1]. To the best of our knowledge, there is no previous work on modeling human behavior in this problem, though it generalizes real-life decision making in various domains like financial investment and security resource allocation.

*This work was done during a PARC internship.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

[1]Note the difference with the multi-armed bandit problem (MAB): In MAB, the DM can only observe the realized sample of *one* random variable (corresponds to the chosen option) per trial and thus faces the tradeoff between exploration and exploitation.

To examine a wide range of candidate human behavior models for the environment learning problem, we define a "model space" by decomposing a human DM's cognitive reasoning process in each trial into two components, i.e the *inference* component and the *selection* component. The inference component describes how DMs aggregate historic information and make forecast for the current trial, while the selection component models how DMs compare different options based on the forecasts and make decisions. Such decomposition provides us with the possibility to conduct a scalable search through the vast space of decision-making models. Moreover, inference and selection models that we evaluate in this paper vary in their assumptions on rationality and thus enable us to pinpoint a few key irrationalities that lead to suboptimal human behavior. As a complementary, we also investigate another family of models, which instead assumes that human DMs simply follow different rules of thumb *without* actively predicting future outcomes in preparation for evaluating alternatives.

From an agent designer's perspective, however, understanding a particular human DM's behavior in a specific decision-making environment is not the whole story. In fact, human DMs' behavior in real life can be influenced by characteristics of the decision-making environment like the degree of uncertainty in the environment [3, 1], as well as DMs' own characteristics. Both of these variations pose new challenges for virtual agent design, as they may imply the needs for creating a huge number of different agents that are tailored to all kinds of contexts and individuals. To better leverage our understanding on human behavior for agent design, we further ask two questions: First, if the goal is to design an agent which resembles an average human DM, does a single type of model exist which is robust against different environment? Second, when creating multiple agents to reproduce the behavior of a population of heterogeneous human beings, how many agents can still be described by an average DM's behavior model and how should we adjust the degree of heterogeneity in the agent population according to the environment?

To this end, we design and conduct a large-scale randomized online experiment, through which we collect data on 24 sequential decisions in an environment learning problem from each of 800 unique human subjects on Amazon Mechanical Turk (MTurk). Each subject of our experiment is randomly assigned to one of the eight treatments with different predefined decision-making environment characteristics. Performance of various computational models is then examined for the "average subject" in each treatment (by assuming an one-size-fit-all model for all subjects in one treatment) and for each individual subject.

Our results suggest that across different decision-making environment, a specific type of two-component model generally outperforms other models, including different rules of thumb, in explaining an average human DM's behavior in the environment learning problem. According to this model, the average DM relies on the most recent information heavily to update their understanding of the uncertain environment, tends to stick with their previous choices, and is inclined to select suboptimal options with higher forecasted utilities. On the other hand, in characterizing each human DM's behavior, we find that while the behavior of a significant portion of individual DMs can be described by the model for the average DM, the degree to which individuals differ from each other is largely affected by the environment. For example, in a less uncertain environment where the sequence of noisy observations implies consistent information, individual DMs are more heterogeneous and display varying levels of rationality; while in a more uncertain environment, they tend to be more homogeneous and behave uniformly naive.

These findings provide valuable implications for designing human-like virtual agents in repeated decision making under uncertainty. On the one hand, the existence of a single type of robust model (against the environment contexts) for resembling average human beings and the fact that this model can also explain the behavior of a large number of individuals indicates a possible decrease of computational burdens for virtual agent design, as we may not need to use a different model to construct an agent every time given a new problem instance. On the other hand, our observations on the relationship between the decision-making environment characteristics and the degree of heterogeneity among human DMs suggest possible principles on tuning behavior in the agent population for context adaptation.

2. RELATED WORK

Various frameworks have been adopted to create virtual agents in uncertain environment, including Fuzzy Logic [9], rule-based or experience-based inference [9, 14], decision network [28], BDI (i.e. belief-desire-intention) and E-BDI (i.e. Emotional BDI) models [15, 13]. In particular, to generate more human-like behavior under uncertainty, [15] integrated the prospect theory [10] with a BDI model and drew a distinction between risk and ambiguity. Our work is different from these studies in two perspectives: First, we focus on a setting where agents have to *repeatedly* make decisions under uncertainty, hence the agent's decision in one trial can be explicitly influenced by not only the current imperfect observation of the environment, but also historic observations and decisions. Second, instead of providing a generic decision-making framework with certain human behavior models embedded a priori, we take a different approach by focusing on a specific decision-making scenario (i.e. the environment learning problem) and examining a variety of different behavior model alternatives in searching for the one which *actually* characterizes real human behavior the best.

Decision-making under uncertainty is a field that is extensively studied in economics. The first, and for a long time the only model in this field, is the expected utility model, which is based on the hypothesis that in an uncertain environment individuals always choose the options that maximize their expected utilities [26]. However, as empirical evidence that violates this hypothesis being consistently observed (e.g. the Allais paradox [2]), new theories have been proposed. One of the alternatives is the random utility model [16], which states that the utility of an option is composed by an observable part (e.g. expected utility) and an unobservable stochastic error term. With different underlying structure in the error terms, various discrete choice models are proposed to account for the choices over multiple options [24]. In our study, while we investigate two selection models that are directly derived from the expected utility and random utility models (i.e. the ϵ-Greedy and the Logit model), given the nature of repeated decision-making in our scenario, we also propose two new selection models (i.e. the Single Hurdle and the Double Hurdle model), with the assumption that subsequent decisions are not independent with each other.

There is an important distinction between these classical decision making under uncertainty settings in economics and the environment learning problem, however. That is, in the environment learning problem, the utility of each option is *unknown* at the decision time[2]. Thus, a decision maker has to *predict* how "good" each option is by inferring from her previous experiences, or equivalently, by forecasting the random variable values based on historic observations. A large number of approaches can be found in statistical inference literatures along this direction, including time-series analysis techniques [7], reinforcement learning methods [22] and the Bayesian inference framework [8]. While most existing virtual agent systems apply the Bayesian inference framework to reason about uncertainty [15, 28], we also consider the other two possibilities and investigate three inference models in this paper (i.e. the Last-K, the TDRL and the Bayesian updating model).

Finally, it has been suggested by behavioral economists and psychologists that people often make decisions based on approximate rules of thumb rather than rational thoughts [25, 23]. In the context of repeated decision-making under uncertainty, a number of rules of thumb have also been specified for other scenarios (e.g. the multi-arm bandit problem) in previous studies [21, 29]. By adapting them into our scenario, we are able to compare the performance of two families of models: two-component models and rules of thumb.

3. EXPERIMENTAL DESIGN

We first introduce our experimental design. Our study is based on the human subjects' choice data that we collect in this experiment on an environment learning problem.

To begin with, we now formally define the environment learning problem: Each of the N random variables in the environment is denoted as X_i, which follows a stationary distribution. The initial observation on X_i is x_i^0. In each trial t ($t \geq 1$), the DM chooses among M options. When the chosen option is $Y_t = j$, the DM's utility in trial t is $U_t = f_j(x_{1:N}^t)$[3], where $f_j(\cdot)$ is the payoff function of option j, and x_i^t is the value of X_i in trial t which will be realized *after* the DM's choice. Both the initial observation and the realized values of X_i in each trial are generated by drawing random samples from its distribution independently. The DM's objective is to maximize her cumulative utility $\sum_{t=1}^{T} U_t$ in all T trials.

3.1 Task

Since our goal is to understand the actual human behavior, in our experiment, we are interested in creating a realistic setting of repeated decision making under uncertainty for the subjects. Thus, we frame the environment learning problem as a game on periodically reviewing the electricity bills and choosing the preferred tariff scheme, which is a common practice in daily life. Specifically, each subject is told that the energy company of her community begins to provide three (i.e. $M = 3$) different electricity tariff schemes:

- *Flat-rate scheme*: Unit price is \$0.25/kWh for electricity usage at any time in a day.
- *Cheaper in the day scheme*: Unit price is \$0.20/kWh during the daytime (7am–7pm) and \$0.30/kWh during the nighttime (7pm–7am).
- *Cheaper in the night scheme*: Unit price is \$0.30/kWh during the daytime and \$0.20/kWh during the nighttime.

One trial in the game corresponds to one "month". In each trial, the subject can review the "electricity bill" of the last month, which includes information on the daytime (nighttime) usage and costs, and choose her preferred tariff scheme for the current month. The game lasts for $T = 24$ trials (i.e. 2 "years"). Figure 1 shows the interface of one trial.

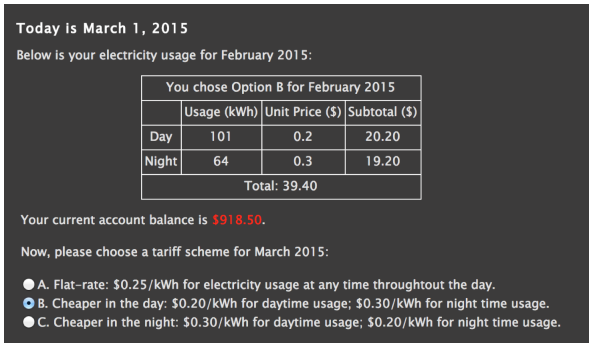


Figure 1: Interface of one trial in the game.

3.2 Treatments

Two random variables (i.e. $N = 2$) are used to represent the daytime (X_1) and nighttime (X_2) electricity usage in each month. We further assume that the electricity usage follows a normal distribution, that is, $X_i \sim N(\mu_i, \sigma_i^2), i \in \{1, 2\}$. To simulate a rich variety of decision-making environment, we control 3 "environment characteristics" to create eight treatments. These characteristics include:

- *Mean usage difference*: the difference between the expected daytime and nighttime electricity usage in a month (i.e. $\Delta\mu = |\mu_2 - \mu_1|$) can be either small or large.
- *Usage variance*: the variance of daytime and nighttime electricity usage (i.e. σ_i^2) can be either small or large.
- *Default option*: when making decisions in each trial, there may or may not exist a default option.

The first two characteristics give us the flexibility to vary the easiness for people to distinguish the underlying distributions of the daytime and nighttime electricity usage, which further influence people's perception on the degree of uncertainty in the environment. The concept of *discriminability*, defined as $d = \frac{|\mu_2 - \mu_1|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, formally captures this idea [1]: With a larger value of $\Delta\mu$ or a smaller value of σ_i^2 (hence larger d), the sequence of noisy samples are more likely to be consistent (e.g. $x_1^t > x_2^t$ in most trials or vice versa) and it's easier for the subjects to figure out their electricity usage is generally higher in one time interval than that in another interval, thus the subjects may perceive the environment to be less uncertain. Different values of these

[2]In classical settings, an option is often presented as a "prospect" (o_1, p_1; ...; o_n, p_n), for which the outcome o_i is yielded with probability p_i and $\sum_{i=1}^{n} p_i = 1$. While the value for p_i may be either known (risk) or unknown (uncertainty), a decision maker is usually aware of the utility of each possible outcome, $u(o_i)$.

[3]The colon notation is used to refer to a range of elements, e.g. $x_{1:N}^t = x_1^t, ..., x_i^t, ..., x_N^t$. In addition, the assumption of a DM's utility being linearly correlated with option payoffs implies that DMs are risk-neutral. Concave/convex utility functions are also tested to account for other risk preferences and results are similar.

two characteristics lead to 4 *electricity usage conditions* and parameter values in each condition are specified in Table 1. Note that within one condition, whether a subject's daytime usage is generally higher (i.e. whether $\mu_1 > \mu_2$) is randomly decided.

Moreover, we include the third characteristic to examine whether human subjects' decisions are affected by the explicit existence of the default (or status-quo) option. A prevalent observation in previous studies is that individuals disproportionately stick with the status-quo option when it exists, which is referred to as the *status-quo bias* [20]. In our experiment, we test 2 *default option conditions*: In the "no default" condition, pre-selected option doesn't exist and an active choice is required for each trial; in the "default" condition, starting from the second trial, the chosen option in the last trial will be pre-selected as the default, yet the subject can choose to switch to other options.

Finally, the combination of an electricity usage condition and a default option condition creates one treatment. Thus, there are 8 treatments in our experiment.

Table 1: Parameter values for the 4 electricity usage conditions.

Conditions	Distributions of X_i	d
Small difference, small variance	$N(70, 10^2)$; $N(90, 10^2)$	1.41
Small difference, large variance	$N(70, 20^2)$; $N(90, 20^2)$	0.71
Large difference, small variance	$N(60, 10^2)$; $N(100, 10^2)$	2.83
Large difference, large variance	$N(60, 30^2)$; $N(100, 30^2)$	0.94

3.3 Experimental Control

Upon arrival, a subject is randomly assigned to one treatment. While the subject understands there may exist some regularities for her electricity usage, she is unaware of the exact forms of the electricity usage distributions in the assigned treatment. In the game, the subject decides only her monthly tariff scheme but *not* the amount of daytime (nighttime) electricity usage shown on the bills, which is actually generated by drawing random samples from the distributions of X_1 (X_2) independently, and the subject has been trained to read the bills. To simulate the real world in which people can benefit from smartly-chosen tariff schemes due to the electricity cost savings, we introduce performance-contingent bonuses in our game. Specifically, Each subject is initially endowed with 1000 dollars of game money and she can observe her account balance in each trial. At the end of the game, besides a fixed payment of \$0.15, the subject may earn a bonus up to \$0.70 by converting the final balance (if positive) with a 100:1 rate. We recommend subjects to keep close track of their monthly bills so as to find out possible regularities hence minimize the costs.

3.4 Data

In total, 800 unique U.S. subjects are recruited from MTurk and 100 subjects are assigned to each treatment, with half of the subjects generally using more electricity in the day (i.e. $\mu_1 > \mu_2$) and the other half consuming more at night (i.e. $\mu_1 < \mu_2$). For each subject, we record the history of daytime (nighttime) electricity usage that she observes (i.e. $x_i^{0:T}$) and her decisions in each trial (i.e. $Y_{1:T}$).

4. MODELS

We now list the formal definitions of all computational models that we examine. In all models, the payoff function of an option j is $f_j(x_{1:N}^t) = -\sum_{i=1}^N w_i^j x_i^t$, where w_i^j is the unit price of daytime or nighttime usage stated in option j[4].

4.1 Two-Component Models: Inference + Selection

When a DM makes decision in trial t, she has observed the realized values of *each* X_i for *all* previous trials (i.e. $x_i^{0:t-1}$), yet her utility in the current trial depends on the unknown value of x_i^t which will be revealed in the future. We hence define our first family of models in a "model space" such that each model consists of two separate components to enable a scalable search over all kinds of decision-making models: the *inference* component is represented by a function that maps the DM's previous observations to forecasts on the random variable values in the current trial, i.e. $\hat{x}_i^t = h(x_i^{0:t-1})$, where $\hat{x}_i^t$ is the predicted value of X_i in trial t; the *selection* component refers to the DM calculating the forecasted utility of each option $\hat{u}_j^t = f_j(\hat{x}_{1:N}^t)$ and stochastically making a choice, i.e. $r_j^t = g(\hat{u}_{1:M}^t)$, r_j^t is the probability of choosing option j in trial t and $\sum_{j=1}^M r_j^t = 1$.

4.1.1 Inference Models

For the inference component, we consider 3 models:

DEFINITION 1 (LAST-K). *The predicted value of random variable X_i in trial t is the average of its realized values in a sliding window of (at most) last K trials: $\hat{x}_i^t = \frac{1}{t-s}\sum_{k=s}^{t-1} x_i^k$, where $s = max(0, t-K)$.* ■

The Last-K model is commonly used in time series analysis (typically referred to as moving average). In addition, an important cognitive foundation of the Last-K model is that humans have limited memory. The widely-known finding of "the magical number 7" suggests that the working memory capacity of an average human is 7±2 [17]. In the limits, when $K = 1$, we have a "*recency model*" in which a DM can only remember the most recent observation; and when $K = t$, we get an "*average model*" that a DM recalls everything from the very beginning.

DEFINITION 2 (TDRL). *The temporal difference reinforcement learning (TDRL) model suggests that a DM predicts the value of a random variable X_i in trial t according to $\hat{x}_i^t = \hat{x}_i^{t-1} + \alpha(x_i^{t-1} - \hat{x}_i^{t-1})$, where α is a learning rate parameter.* ■

TDRL is a prediction method mostly used for solving reinforcement learning problems [22, 6]. The core idea is that in each trial, the prediction is adjusted to better match the most recent observation, with a larger α indicating a higher discount on the old information.

Both the Last-K and the TDRL model imply that DMs may have bounded rationality so they apply heuristics like trimming and weighting when aggregating information and making inference. On the contrary, Bayesian inference, which is the framework frequently used in existing virtual agent systems, presents a idealistic inference mechanism by assuming that DMs have sufficient prior knowledge (e.g. the

[4]The negative sign converts a cost minimizing problem into a utility maximizing problem.

models that generate the observations) and rationally make inferences by applying the Bayesian updating scheme given the sequential observations.

DEFINITION 3 (BAYESIAN UPDATING). *In each trial, a DM updates her posterior belief on parameter values for the distribution of X_i by integrating the realized value of X_i with the prior belief using the Bayes rule, and her prediction $\hat{x}_i^t$ for trial t is the expected value of X_i conditioned on its posterior parameter distribution in trial $t-1$.* ■

For our case, a DM's initial belief on the two parameters for the distribution of X_i (i.e. μ_i and the precision parameter $\lambda_i = 1/\sigma_i^2$) is assumed to be a Normal-Gamma distribution as the conjugate prior[5].

4.1.2 Selection Models

For the selection component, we study 4 different models:

DEFINITION 4 (ϵ-GREEDY). *A DM chooses the (set of) optimal option(s) with probability ϵ:*

$$r_j^t = \begin{cases} \frac{\epsilon}{|J^*|} & j \in J^* \\ \frac{1-\epsilon}{M-|J^*|} & j \notin J^* \end{cases}$$

where J^* is the set of options with the maximal forecasted utility $\hat{u}_j^t$ in trial t. ■

The ϵ-Greedy model corresponds to a stochastic version of the expected utility model by stating that DMs will choose the option with maximal utility, yet there is a constant chance of making errors (i.e. $1-\epsilon$).

DEFINITION 5 (LOGIT). *The probability for a DM to choose option j is given by $r_j^t = \frac{exp(\beta \hat{u}_j^t)}{\sum_{j'=1}^{M} exp(\beta \hat{u}_{j'}^t)}$, where β is a precision parameter.* ■

The Logit model is a prominent type of discrete choice model in economics [24], which is based on a random utility model assuming that the stochastic error term in utility follows a Gumbel distribution (i.e. type I generalized extreme value distribution). It uses a softmax function to convert option utility into choice probability. The parameter β indicates a DM's sensitivity to utilities: when $\beta \to 0$, the DM chooses randomly; when $\beta \to \infty$, the DM almost always chooses the forecasted optimal option. On the cognitive level, the Logit model implies a property called *cost-proportional errors*, that is, human DMs can be more likely to err when the error costs are lower (i.e. choose suboptimal options more often when they have larger utilities), and this property has also been modeled as the *quantal best response* in behavior theory [27].

Both the ϵ-Greedy and the Logit model assume that human DMs evaluate options in each trial independently. However, when a DM makes decisions repeatedly, her choice in a specific trial may be affected by her previous decisions, for example, she may be inclined to keep her previous choice. Inspired by the hurdle and zero-inflated model in statistics, which are developed to characterize count data with extra zero-valued observations [18, 11], we propose the following two models to take such limitation into account:

[5]That is, $NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) = N(\mu|\mu_0, (\kappa_0\lambda)^{-1}) \cdot Ga(\lambda|\alpha_0, \beta_0)$, where $N(\cdot)$ and $Ga(\cdot)$ are the probability density functions of Normal and Gamma distributions, respectively.

DEFINITION 6 (SINGLE HURDLE). *When $t = 1$, the Single Hurdle model is the same as the Logit model. Otherwise, the probability for a DM to choose option j is given by:*

$$r_j^t = \begin{cases} \pi & j = Y_{t-1} \\ \frac{(1-\pi)exp(\beta \hat{u}_j^t)}{\sum_{j' \neq Y_{t-1}} exp(\beta \hat{u}_{j'}^t)} & j \neq Y_{t-1} \end{cases}$$

where β is a precision parameter and Y_{t-1} is the DM's choice in trial $t-1$. ■

DEFINITION 7 (DOUBLE HURDLE). *When $t = 1$, the Double Hurdle model is the same as the Logit model. Otherwise, the probability for a DM to choose option j is given by:*

$$r_j^t = \begin{cases} \pi + \frac{(1-\pi)exp(\beta \hat{u}_j^t)}{\sum_{j'=1}^{M} exp(\beta \hat{u}_{j'}^t)} & j = Y_{t-1} \\ \frac{(1-\pi)exp(\beta \hat{u}_j^t)}{\sum_{j'=1}^{M} exp(\beta \hat{u}_{j'}^t)} & j \neq Y_{t-1} \end{cases}$$

where β is a precision parameter and Y_{t-1} is the DM's choice in trial $t-1$. ■

These two models implicitly assume 2 steps in selecting options. In Step 1, the DM decides whether to stick with her choice in the last trial regardless of its forecasted utility, and this is governed by a binomial distribution with parameter π. In Step 2, the DM makes further evaluations if the previous choice is not preserved in Step 1: For the Double Hurdle model, *all* options are evaluated; yet for the Single Hurdle model, the previous choice is *excluded* from the evaluation. The conditional probability of selecting an option in Step 2 follows the Logit model.

4.2 Rules of Thumb

The second family of models is a few rules of thumb, assuming that human DMs don't explicitly predict future outcomes and make decisions based on the forecasted utility of each option. Four rules are examined in this study:

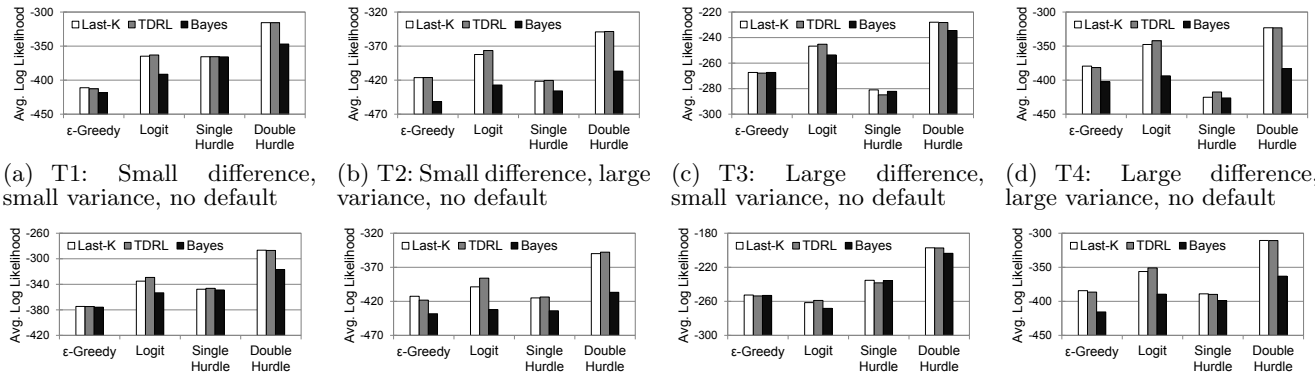
RULE 1 (RANDOM). *A DM chooses an option uniformly randomly in each trial. That is, option j is selected in trial t with probability $\frac{1}{M}$.* ■

The random rule suggests a naive way of making decisions in regardless of utilities. Thus, it is a baseline rule of thumb.

RULE 2 (PROBABILITY MATCHING). *A DM initially assigns a "success count" n_j^0 to each option j. Upon the observation of $x_{1:N}^{t-1}$ before trial t, if $f_j(x_{1:N}^{t-1}) \geq f_{j'}(x_{1:N}^{t-1})$ for any j', the DM updates the success count of option j as $n_j^t = n_j^{t-1}+1$; otherwise $n_j^t = n_j^{t-1}$. Option j is then selected in trial t with probability $\frac{n_j^t}{\sum_{j'=1}^{M} n_{j'}^t}$.* ■

Probability matching suggests a suboptimal decision strategy that has been widely observed among humans: Instead of consistently choosing the maximizing option, the DM assigns a probability to each option by matching its likelihood of being optimal [21].

RULE 3 (GOOD-STAY-BAD-SHIFT). *Upon the observation of $x_{1:N}^{t-1}$ before trial t, if the DM's previous choice Y_{t-1} turns out to be "good", the DM will keep that choice with probability p; otherwise, the DM will shift to the (set of) optimal option(s) (according to $x_{1:N}^{t-1}$) with probability p. Other options are selected uniformly randomly in both cases.* ■



(a) T1: Small difference, small variance, no default

(b) T2: Small difference, large variance, no default

(c) T3: Large difference, small variance, no default

(d) T4: Large difference, large variance, no default

(e) T5: Small difference, small variance, default

(f) T6: Small difference, large variance, default

(g) T7: Large difference, small variance, default

(h) T8: Large difference, large variance, default

Figure 2: Mean log likelihood in 5-fold cross validation when modeling an average DM using two-component models.

We test two ways to define a "good" option: an aggressive DM requires Y_{t-1} to have the highest utility among all options given $x_{1:N}^{t-1}$, yet a conservative DM considers Y_{t-1} to be good as long as its utility is not the lowest. This rule is adapted from the "win-stay-lose-shift" heuristics in previous studies for the two-armed bandit problem [29].

Finally, we consider a fourth rule which describes the DM as constantly choosing a "safe choice" unless another option can be significantly better.

RULE 4 (SAFE CHOICE). *A DM has a safe choice j_s in mind. Upon the observation of $x_{1:N}^{t-1}$, the DM will only switch to the optimal option j^* (according to $x_{1:N}^{t-1}$) in trial t with probability q if $\frac{f_{j^*}(x_{1:N}^{t-1}) - f_{j_s}(x_{1:N}^{t-1})}{|max(f_{j^*}(x_{1:N}^{t-1}), f_{j_s}(x_{1:N}^{t-1}))|} > \theta$; otherwise, the DM will stay with her safe choice with probability q. Other options are selected uniformly randomly in both cases.* ■

A natural safe choice in our setting is the "flat-rate scheme" as it can't be the worst option at any time. Such preference for the flat-rate scheme is also observed empirically, which may because humans tend to avoid variations [12].

5. RESULTS

In this section, we present the performance of each computational model in explaining human DMs' behavior, both for an average DM and for each individual DM.

5.1 Modeling for an Average DM

Our first research question is how to design an agent to resemble an *average* human DM in repeated decision making under uncertainty. Therefore, we start with the evaluation on the performance of different models (including both two-component models and rules of thumb) in terms of explaining the average human DM's behavior in various decision-making environment.

To evaluate the performance of a specific type of model in a particular decision-making environment (i.e. treatment), we conduct a 5-fold cross validation, and an *one-size-fit-all* model of the given type is used to account for the average human DM's behavior. That is, given the dataset for one treatment (i.e. all choice data for all subjects in that treatment), we first create 5 folds by randomly partitioning the subjects into 5 groups and collecting all data for a subject to her group. Next, we retain one fold of the dataset as the validation dataset while the other 4 folds are used as training data, and the cross-validation process is repeated 5 times by using each fold as the validation dataset once. Within one round of cross validation, given a model type (e.g. Last-K + ϵ-Greedy), we first train an average DM's model of this type using maximum likelihood estimation, with the assumption that all subjects in the training dataset share the same model parameters and thus search for the optimal parameter values which give us the largest probability of observing the training data. Then, we calculate the log likelihood value of the validation data given the obtained model. The mean log likelihood value of all 5 folds is used to represent how well this type of model fits an average DM's behavior in this treatment, with a larger value indicating a better model. Hence, by comparing the mean log likelihood values of different models within each treatment, we can understand what is the best model to characterize the average DM's behavior in each decision-making environment.

5.1.1 The Best Two-Component Model

We first examine the performance of different two-component models in various environment, and the results are shown in Figure 2(a)-2(h). When the Last-K model is used for the inference component, we test on a number of different K values (i.e. $K = 1, 2, 3, 6, 12, t$) and the largest log likelihood value for the corresponding two-component model is reported. As the figures suggest, in *all* eight decision-making environment, the combination of the TDRL (or the best Last-K) inference model and the Double Hurdle selection model consistently has a better performance than other two-component models in capturing an average DM's behavior.

Particularly, for the inference component, it is interesting to see that the widely-used model for virtual agents, that is, the Bayesian updating model, actually describes an average real human DM's behavior relatively poor. This can be possibly attributed to human DMs' limited prior information and rationality. Indeed, a closer look at the other two inference models suggest that the average DM is irrational in the reasoning process: While the TDRL model consistently reports very high learning rates (e.g. estimated α values are in the range of 0.86-0.99), we also find that the recency model (i.e. $K = 1$) almost always provides the highest log likelihood among all Last-K models when combined with a selection model (except in $T3$). These observations indicate that the average human DM is subject to the *recency bias* severely

when she predicts the future based on the history, that is, she reacts strongly to the new information and discounts old information heavily [5, 6].

On the other hand, when we focus on the selection component, one of the models that we propose, i.e. the Double Hurdle model, clearly outperforms other models. The implications are two-folds: First, instead of being a strict utility maximizer, the average DM makes cost-proportional errors (e.g. estimated β values are in the range of 0.72-1.52); Second, when the average DM makes decisions repeatedly, she doesn't take each of them independently. In fact, the average DM is affected by the status-quo bias and is reluctant to change decisions in subsequent trials, even if the previous choice is *not* explicitly set as the default (e.g. estimated π values suggests a 25%-50% chance of sticking with the previous choice). The average DM's bias towards the status-quo option also serves as an add-on to her analysis on the forecasted utility of each option, which leads to the advantage of the Double Hurdle model over the Single Hurdle model.

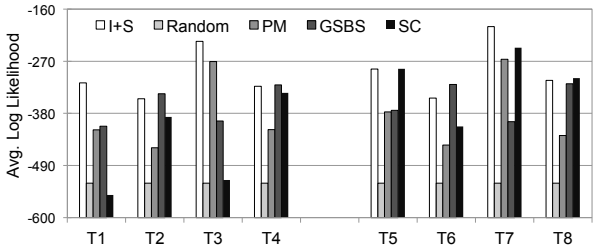


Figure 3: Comparison between the best two-component model and different rules of thumb.

5.1.2 Inference+Selection vs. Rules of Thumb

Figure 3 presents the comparison between the two families of models in fitting an average DM's behavior, where $T1$-$T4$ are treatments without the default option and $T5$-$T8$ are treatments with the default option. When variations exist for a model (e.g. different combinations of inference and selection component for the two-component model, the aggressive and conservative way to define a "good" option in a Good-Stay-Bad-Shift rule, etc.), the one with the highest log likelihood value is reported.

Clearly, the best two-component model fits the data reasonably well for *all* 8 treatments, which essentially suggests the combination of a TDRL inference model and a Double Hurdle selection model can be generally recommended for designing a virtual agent to simulate an average DM in the environment learning problem, no matter what the specific decision-making environment is. In contrast, different rules of thumb may capture the average human DM's behavior accurately in *some* specific environment, for example, the Safe Choice rule performs well in treatments with the default option. However, none of these rules provides robustly good fitting performance across all treatments, indicating that for the average DM, the activation of rules of thumb, if any, is highly context-dependent, which is also consistent with the literatures [19].

5.2 Modeling for Individual DMs

Our second research question is when designing an agent population to reproduce the behavior of a group of heterogeneous human DMs, how many agents in the population can still adopt the best-performing model for the average DM and how should we adjust the degree of heterogeneity within the population based on the decision-making environment. We hence proceed on to characterize *individual* DM's behavior using various two-component models. That is, we treat each individual DM as a unique entity and explore a "personalized" behavior model for every one of them. As subjects don't exhibit significantly different behavior in the 2 default option conditions, in this subsection, we combine subjects together for each of the 4 electricity usage conditions.

5.2.1 Validity of the Average DM's Model on the Individual Level

To answer the question of to what degree the best-performing average DM's behavior model can be used if we aim at creating a population of heterogeneous agents, we seek to find out the fittest inference model and the fittest selection model among individual DMs, respectively.

For the fittest inference model, we first set each DM's selection model to be the Double Hurdle model and compare the fitness of different inference models (i.e. the log likelihood value) in explaining the individual DM's decisions. Figure 4(a) reports the percentage of subjects in each electricity usage condition that the best inference model is Last-K, TDRL and Bayesian updating, respectively. As shown in the figure, across all four different environment, the fittest inference model is the TDRL model for the majority of individual DMs, which is consistent with the inference component choice in the best-performing model for the average DM.

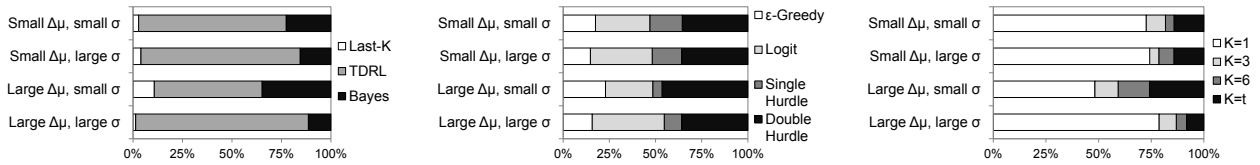
Moreover, for the fittest selection model among individuals, we fix a DM's inference model to be the TDRL model and compare the performance of different selection models for each individual. Results are presented in Figure 4(b). While we observe more diversity in DM's choice on the selection model, we still find the Double Hurdle model to be the fittest selection model for the largest percentage of individual DMs. That is, the selection model in the best-performing model for the average DM can also be used for a large portion of individual DMs.

In sum, through the examination on the fittest inference and selection model for each individual DM, we validate the feasibility of the best-performing average DM's behavior model on the individual level by showing that it can actually be used to characterize the behavior of a large number of individual DMs.

5.2.2 Heterogeneity and the Environment

The existence of individual differences in a heterogeneous population indicates that there are always some individuals who "deviate" from the average. For example, as Figure 4(a) shows, depending on the decision-making environment, 11%-35% of individual DMs actually display consistency with the Bayesian inference framework, suggesting a subgroup of individuals with possibly higher levels of rationality. We hence ask how is the degree of heterogeneity among individual DMs affected by the decision-making environment.

Take the heterogeneity among individuals on making inferences about the uncertain environment as an example. Figure 4(a) seem to indicate that DMs are more heterogeneous in the "Large difference, small variance" condition, as more individuals are able to avoid the recency bias hence behave more rationally. More formally, the degree of het-



(a) The inference component (b) The selection component (c) The Last-K model

Figure 4: Individual DM's behavior in the environment learning problem.

erogeneity within a population can be measured by Shannon's entropy. In particular, given a population of L types of individuals, if the probability of an individual belonging to type l is p_l, the heterogenous degree of this population can be defined as $H = -\sum_{l=1}^{L} p_l log(p_l)$, with a larger value representing a higher degree of heterogeneity. The heterogenous degree for each population in the four decision-making environment in Figure 4(a) is reported in the H_{inf} column of Table 2. As we expect, the values differ a lot across various environment, while individuals in the "Large difference, small variance" environment have the highest degree of heterogeneity.

Recall from Table 1 that the "Large difference, small variance" environment has the highest discriminability in all 4 conditions, implying that the sequence of observations are likely to provide consistent information and DMs may perceive this environment to be less uncertain. Thus, we make a conjecture on the relationship between the decision-making environment and the degree of heterogeneity among human DMs: In an environment of high discriminability (i.e. less perceived uncertainty), individuals are capable of leveraging their varying levels of rationality hence are more heterogeneous; however, in an environment of low discriminability (i.e. more perceived uncertainty), individuals uniformly display their irrationality hence are more homogeneous.

As another supporting evidence, we set each DM's inference and selection model to be the Last-K and Double Hurdle model, respectively. We then test different K values (i.e. $K = 1, 3, 6, t$) and rank the fitness of each model for each DM. Figure 4(c) shows the percentage of subjects for whom the best K value is $1, 3, 6, t$, respectively, when the Last-K model is used as the inference component, and the heterogenous degree of each population in this figure is also reported in the H_{Last-K} column of Table 2. Again, we find that in a less uncertain environment, DMs are more heterogeneous and more individuals consider a longer sequence of previous observations, or even the entire observation history, to predict the future. Yet, in a more uncertain environment, most of DMs display recency bias and as a result, the population is more homogeneous.

Table 2: The degree of heterogeneity among individual DMs on making inferences for the 4 electricity usage conditions.

Conditions	H_{inf}	H_{Last-K}
Small difference, small variance	0.6558	0.8586
Small difference, large variance	0.5925	0.8246
Large difference, small variance	0.9378	1.2282
Large difference, large variance	0.4269	0.7453

6. DISCUSSIONS AND CONCLUSIONS

In this paper, we evaluate and compare the performance of a variety of computational models in fitting the data of real human behavior in an environment learning problem. Based on our evaluation results, we list two guidelines for designing human-like virtual agents in repeated decision making under uncertainty:

1. To resemble an average person, the agent should be designed with a few human irrationalities, like the recency and status-quo bias and tendency to make cost-proportional errors.
2. While the same behavior model can be used across different decision-making environment for agents that resemble average human beings, when designing an agent population to simulate a group of heterogenous people, we should adjust the degree of heterogeneity within the population (i.e. the use of different models for each individual agent) according to the decision-making environment.

The first guideline advocates for deeper considerations of the cognitive limitations and biases of human beings for the design of human-like virtual agents. Essentially, the question is whether a fully rational agent is really "human-like", and we think the answer is "No".

The second guideline is based on our findings that the combination of the TDRL inference model and the Double Hurdle selection model provides accurate description of an average DM's behavior across all decision-making environment, yet the specific environment characteristics largely influence how much individual DMs differ from each other. Particularly, we observe that people seem to be more heterogeneous in a less uncertain environment while more homogeneous in a more uncertain environment.

There are many interesting work that can be done in the future. For example, in this paper, we provide a methodology to conduct scalable search through the vast space of human behavior models in repeated decision making under uncertainty by decomposing the decision-making process into two components. While we strive to cover a wide range of both inference and selection models in the study, our search may not be comprehensive. More sophisticated models, such as various complicated time-series and discrete choice models, can be integrated into the space and be evaluated in the future. Furthermore, designing virtual agents which have these human behavior models in mind and target to interactively train or "nudge" real people towards better decisions would be another very exciting future direction.

Acknowledgements

The authors are grateful to Yiling Chen and Barbara Grosz for helpful discussions and feedback. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

[1] H.-i. Ahn and R. W. Picard. Modeling subjective experience-based learning under uncertainty and frames. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[2] M. Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546, 1953.

[3] A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio. Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295, 1997.

[4] T. Bosse, C. Gerritsen, and J. Treur. Combining rational and biological factors in virtual agent decision making. *Applied intelligence*, 34(1):87–101, 2011.

[5] C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

[6] D. Fudenberg and A. Peysakhovich. Recency, records and recaps: learning and non-equilibrium behavior in a simple decision problem. In *Proceedings of the fifteenth ACM conference on Economics and Computation*, pages 971–986. ACM, 2014.

[7] J. D. Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[8] C. Howson and P. Urbach. *Scientific reasoning: the Bayesian approach*. Open Court Publishing, 2006.

[9] I. Hupont, R. Del-Hoyo, S. Baldassarri, E. Cerezo, F. J. Serón, and D. Romero. Towards an intelligent affective multimodal virtual agent for uncertain environments. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, page 4. ACM, 2009.

[10] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[11] D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[12] A. Lambrecht and B. Skiera. Paying too much and being happy about it: Existence, causes, and consequences of tariff-choice biases. *Journal of Marketing Research*, 43(2):212–223, 2006.

[13] H. Lejmi-Riahi, F. Kebair, and L. B. Said. Agent decision-making under uncertainty: Towards a new e-bdi agent architecture based on immediate and expected emotions.

[14] L. Luo, S. Zhou, W. Cai, M. Lees, and M. Y.-H. Low. Modeling human-like decision making for virtual agents in time-critical situations. In *Cyberworlds (CW), 2010 International Conference on*, pages 360–367. IEEE, 2010.

[15] N. Marques, F. Melo, S. Mascarenhas, J. Dias, R. Prada, and A. Paiva. Towards agents with human-like decisions under uncertainty. In *The Annual Meeting of the Cognitive Science Society*, 2013.

[16] D. McFadden. Economic choices. *American Economic Review*, pages 351–378, 2001.

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

[18] J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

[19] J. W. Payne, J. R. Bettman, and E. J. Johnson. *The adaptive decision maker*. Cambridge University Press, 1993.

[20] W. Samuelson and R. Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, 1988.

[21] L. P. Sugrue, G. S. Corrado, and W. T. Newsome. Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787, 2004.

[22] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

[23] R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, 2008.

[24] K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge university press, 2009.

[25] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[26] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior (2nd rev). 1947.

[27] J. R. Wright and K. Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. In *AAAI*, 2010.

[28] Q. Yu and D. Terzopoulos. A decision network framework for the behavioral animation of virtual humans. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 119–128. Eurographics Association, 2007.

[29] S. Zhang and J. Y. Angela. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in Neural Information Processing Systems*, pages 2607–2615, 2013.