Evaluating the Applicability of Peer-Designed Agents in Mechanisms Evaluation

Avshalom Elmalech
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
elmalea@cs.biu.ac.il

David Sarne
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
sarned@cs.biu.ac.il

***Abstract*—In this paper we empirically investigate the feasibility of using peer-designed agents (PDAs) instead of people for the purpose of mechanism evaluation. This latter approach has been increasingly advocated in agent research in recent years, mainly due to its many benefits in terms of time and cost. Our experiments compare the behavior of 31 PDAs and 150 people in a legacy eCommerce-based price-exploration setting, using different price-setting mechanisms and different performance measures. The results show a varying level of similarity between the aggregate behavior obtained when using people and when using PDAs — in some settings similar results were obtained, in others the use of PDAs rather than people yields substantial differences. This suggests that the ability to generalize results from one successful implementation of PDA-based systems to another, regarding the use of PDAs as a substitute to people in systems evaluation, is quite limited. The decision to prefer PDAs for mechanism evaluation is therefore setting dependent and the applicability of the approach must be re-evaluated whenever switching to a new setting or using a different measure. Furthermore, we show that even in settings where the aggregate behavior is found to be similar, the individual strategies used by agents in each group highly vary.**

***Keywords*-PDA; system evaluation;**

I. Introduction

Recent research increasingly relies on Peer-Designed Agents (PDAs), computer agents developed by people, as an efficient means for replacing people in evaluating systems [21]. Examples of such implementations can be found in various domains, such as negotiation [16], security systems [15] and parking allocation [3]. The use of agents in general for system evaluation encompasses many advantages, such as allowing the simulation of high-level types of information-oriented behaviors, interaction with other individuals, influencing one another to make separate decisions, and simulation of large-scale systems, due to the relatively small cost of cloning agents [31]. The key challenge in using agents to replace people for system evaluation is in having the ability to accurately capture people's behavior. This is mostly because people are known to be associated with diverse behaviors, which makes it difficult to capture behavior patterns in a monolithic model [17]. The use of PDAs in comparison to other agent design methods may offer a better representation of the rich set of behaviors used by people.

The underlying assumption in many of the PDA-based works is that PDAs capture people's behavior adequately, and therefore a PDA-based system is likely to perform similar to the case where the system is populated with people [4]. Empirical investigation of the level of similarity observed between PDAs and people is not conclusive; some work suggests a relatively strong correlation between the behaviors of the two [4], while in other work the PDAs are reported to act different than people to some extent [8], [21]. Yet, even in cases where individual differences between the behavior of PDAs and people were reported, the performance of mechanisms applied directly on people and PDAs were similar [16], [15]. Consequently, the conclusion of those works is that PDAs can alleviate the evaluation process of mechanisms (replacing people) and facilitate their design.

Unlike prior work, the underlying hypothesis of the research reported in this paper is that the extent of similarity between the average behavior observed when using people and PDAs is mechanism and measure dependent. The reported experimental results demonstrate how the use of the same set of PDAs leads to various different conclusions regarding the applicability of the "PDAs as a substitute for people" approach, simply by changing either the measures used or the mechanism evaluated. While for some combinations a complete similarity is reported, in others the difference is substantial.

In order to test this hypothesis, a set of experiments with people and PDAs is carried out. The experimental design uses the classical exploration-versus-exploitation problem of price-search in eCommerce [29]. The evaluation of the the performance of people and PDAs is measured by three different parameters (measures): the exploration extent, the agent's overall expense and the seller's revenue. The evaluation of performance recurs with three different price setting mechanisms, using 150 human subjects and 31 PDAs.

The importance of the results is in showing that great caution should be used when attempting to generalize specific results obtained in this line of research — one cannot deduce from the success of PDAs in reflecting an average behavior similar to the one obtained with people in a specific domain to other domains. Furthermore, even if using the same domain, there is no guarantee that the same level of similarity holds when changing the evaluated mechanism. Even if evaluating just one mechanism, but changing the performance measure, there is no guarantee that similar results would be obtained. Instead, whenever considering the use of PDAs as a substitute for people in system evaluation, one needs to test the applicability of this approach for the specific mechanism and measures of interest, using a pilot study or other means.

In the following section we review related work. Then we introduce the exploration problem that is used in this paper as a framework for evaluation, the experimental design and results analysis. We conclude with a discussion concerning the implications of the results.

II. Related Work

The use of PDAs in multi-agent system literature is quite extensive. For example, in Kasbah [5] PDAs that buy and sell are used for evaluating an electronic marketplace. In Colored Trails [8], PDAs are used for reasoning about players' personalities in uncertain environments. Other works, e.g., [16], [15], [3], [21] used PDAs for evaluating specific mechanisms in various domains such as evaluating security algorithms, evaluating automated negotiators, etc. This concept of requesting people to program strategies into agents emerged from behavior economics, where the "strategy method" paradigm, according to which people state their action for every possible situation that may arise, is widely in use [26], [25], [19].

The main motivation for having people program complex agents' behaviors is the possibility that the resulting strategy will correspond to their own. This, however, is not straightforward. Evidence of discrepancies between actual and reported human behavior is a prevalent theme in research originating in various other domains, in particular in metacognition research [9]. Examples of such discrepancies include over-reporting of political participation (roughly 25% of non-voters report of having voted immediately after an election [2]) and contrasting results between self-reported and performance-based levels of physical limitations (there is weak to moderate association between performance-based and self-reported measures in motor functioning [13]). Part of the PDA-based literature simply does not consider the PDAs-people similarity question as an issue or attempt to make any claims regarding this aspect (e.g., in TAC [30]). Yet, much of the PDA literature tends to assume that people can successfully (to some extent) capture their real-life strategy in a given domain when programming an agent. Encouraging results concerning the relatively high similarity between people and PDAs are reported in [4]. Even in cases where some discrepancy between PDAs and people's behavior is reported, the average performance is reported to be similar, suggesting that PDAs can replace people in mechanism evaluation [16], [15]. Common to all the above work is that they draw their people-PDAs similarity-related conclusions based on the specific implementation they use. The current work attempts to present a more targeted evaluation, which compares the same populations while varying both the mechanism and measure used.

III. The Exploration Model

As the underlying framework for the research, we consider the canonical sequential exploration problem described by Weitzman [29] to which a broad class of search problems can be mapped. In this problem, a searcher is given a number of possible available opportunities $S = \{s_1, ..., s_n\}$ (e.g., to buy a product) out of which she can choose only one. The value v_i to the searcher of each opportunity s_i (e.g., expense, reward, utility) is unknown. Only its probability distribution function, denoted $f_i(v)$, is known to the searcher. The true value v_i of opportunity s_i can be obtained, but only by paying a fee, denoted c_i, possibly different for each opportunity. Once the searcher decides to terminate her search (or once she has uncovered the value of all opportunities) she chooses the one with the minimum or maximum value (depending on whether values represent costs or benefits) among the opportunities whose values were obtained.

The exploration problem as formulated above is generic and applies to a variety of real-world situations. For example, when looking for a used car, the information given in ads reflects a noisy signal, and the true value of cars can be revealed only through a test drive or a costly inspection. Similar exploration problem characteristics can be found in job-search [7], multi-robot systems [10] and assignment of jobs to servers in computer systems [22].

The optimal (overall-expense minimizing) strategy for the above exploration problem is given in "search theory" literature [29]. It involves setting a threshold, denoted r_i, for each opportunity (e.g., a seller) s_i, calculated according to:

$$c_i = \int_{x=-\infty}^{r_i} (r_i - x) f_i(x) dx \tag{1}$$

Intuitively, r_i is the value where the agent is precisely indifferent: the expected marginal benefit from obtaining the value of the opportunity exactly equals the cost of obtaining that additional value. The agent should always choose to obtain the value of the opportunity associated with the minimum threshold and terminate the exploration once the minimum value obtained so far is less than the minimum threshold of any of the remaining opportunities.

While this exploration setting is common, the nature of its optimal strategy is non-intuitive therefore, people's exploration strategies when facing the problem are diverse and likely to be different from the optimal strategy. For example, an important property of the above solution is that the threshold calculated for each opportunity does not depend on the number and properties of the other opportunities, but rather merely on the distribution of the value of the specific opportunity and the cost of evaluating it.

The specific sequential exploration problem that is used as a framework for testing our hypothesis considers buyer agents operating in an electronic marketplace populated by N seller agents. The sellers are assumed to be active in various other markets (i.e., multichannel retailers) and frequently change their prices through various dynamic pricing techniques [12], [14]. Their pricing is therefore taken to be external, unaffected by the buyer's exploration behaviors in this market. Any seller s_i is therefore assumed to be associated with a different distribution function $f_i(v)$, from which its price is drawn at any given time, if queried for its price.[1] The distributions $f_i(v)$ $(i = 1, ..., N)$ are assumed to be known to the buyer agent (or can be learned using past experience, Bayesian update, etc.). The agent's price-querying process itself is assumed to be costly, as it requires the consumption of some resources (either CPU time, communication bandwidth, etc., in the agents' world or time, parking fees, transportation costs, etc., in the physical world) [14]. This cost is assumed to be seller-dependent — the cost of querying the price of seller s_i is denoted c_i. Buyers are assumed to always prefer the minimum quote among those received from the sellers they query (i.e., sensitive to price [28]).

Given the description above, an agent needs to set its expense-minimizing exploration strategy based on the set of sellers, their distribution of prices and querying costs $(f_1(v), ..., f_N(v)\ , c_1, ..., c_N)$. The agent's strategy thus determines which seller to query next (if at all) based on the best (lowest) price obtained so far and the set of sellers that have not been queried yet.

Reliably capturing the exploration strategies of agents in such settings is important for evaluating various market design mechanisms, among which we focus on price setting mechanisms. In particular, we consider the problem of a single seller that operates only in this market (i.e., a non-multichannel retailer whose pricing does not affect the prices set by the other sellers). A key challenge for a seller whenever considering new price-setting mechanisms (i.e., one that sets a price as a function of the costs of query and distribution of prices of the other sellers) or when having to choose among several existing ones is to reliably evaluate the expected revenue from each such mechanism. Evaluating the mechanisms by actually applying them is risky, as it may be associated with substantial losses. The use of PDAs in this case may seem quite appealing if it can be guaranteed that the resulting system performance (i.e., in terms of average expected revenue) will be similar to the one obtained in the case of evaluating the mechanisms with human buyers.

Overall, the use of the exploration problem described above has many advantages in the context of this research. First, it considers a real-life setting with which most people are familiar (or experienced with, and so are likely to have a well-established strategy for). Second, sellers in such domains are required for frequent evaluations of different price setting methods due to the dynamic nature of eCommerce which makes it ideal for PDA-based evaluation, if indeed PDAs exhibit similar performance (to people). This also implies that various price setting mechanisms can be considered for our experiments. Finally, there are several possible measures for capturing overall exhibited exploration behavior (i.e., overall performance), according to which the similarity between PDAs and people can be evaluated.

IV. Evaluation Methodology

The goals set for the evaluation are twofold: First, we wanted to compare the system's performance when the exploration is carried out by PDAs and people when using different price-setting mechanisms and different measures. Second, we wanted to supply a drill-down analysis of the exploration strategies used by people and PDAs, possibly explaining the differences observed in the system performance. Along with these two goals, we also evaluated the performance of an agent using the optimal (i.e., expected-expense-minimizing) exploration strategy. While the results obtained by this agent do not relate to the main underlying hypothesis of this paper, this agent's performance is a good benchmark for evaluating the extent to which people's and agents' strategies are suboptimal, and how significant the difference between the latter two is.

A. Pricing Mechanisms

The goal of having different pricing mechanisms in our experimental design is to enable producing different sets of exploration settings with a slight, though consistent, variation between sets. With each different pricing method, buyers face the same sets of problems, however the price of one of the sellers is determined according to different logic.

We emphasize that, for the purpose of testing the research hypothesis, any set of mechanisms that acts according to some consistent logic is legitimate. Still, in an effort to improve the realism of our experiments, we attempted to come up with "reasonable" pricing mechanisms, of the kind that sellers who will potentially adopt the PDA-based methodology might choose to evaluate. The three different pricing methods that were designed for our experiments are: Theoretic-Optimal Pricing (M1), Mean-Fonders Pricing (M2) and Cost-Probability Tradeoff Pricing (M3). All three are designed to output a deterministic price q (or alternatively, a degenerated distribution with all of its mass around q). Since the model assumes that the buyer must query the seller at least once prior to making the purchase, then in order to purchase the product at price q the buyer needs to spend $c_i + q$ (c_i is the cost for querying seller i).

All three methods assume that no prior information about the buyer is available and rely solely on the distribution of the other sellers' prices and exploration costs:

1) Theoretic-Optimal Pricing (M1): This price setting method assumes that the buyer is fully rational and uses the optimal exploration strategy described in the former section. In this case the buyer will buy the product from the seller i at price q (with a total expense of $q + c_i$) only if she already queried all of the sellers associated with a threshold smaller than $q + c_i$ and the best price found was above $q + c_i$ (the threshold is calculated for each seller according to Equation 1). The expected benefit from setting a price q is thus given by $q \prod_{r_j < q+c_i} (1 - F_j(q + c_i))$, where $F_j(q)$ is the appropriate cumulative distribution function of the price. The deterministic price q that maximizes seller s_i's expected revenue can therefore be extracted from:

[1] This assumption, which is commonly used in exploration models [29], has evidence in a large body of empirical research in the form of the persistence of price dispersion both in traditional and online retail markets [6], [1].

$$argmax_q\big(q \prod_{r_j<q+c_i}(1-F_j(q+c_i))\big) \quad (2)$$

2) Mean-Fonders Pricing (M2): The Mean-Fonders Pricing is based on experimental evidence relating to the tendency of people to overemphasize mean values in problem solving, reasoning about this one feature of a distribution rather than all of the distribution's features [18]. A buyer whose exploration is driven by mean values is likely to follow a modification of the above threshold-based optimal exploration rule, according to which the thresholds used are merely based on means. Specifically, the threshold that will be assigned to each seller s_i in this case is $r'_i = \mu_i + c_i$, where μ_i is the expectancy of $f_i(q)$.

The buyer (i), following the mean-based exploration strategy will thus buy the product from the seller using a deterministic price q and an exploration cost c_i, only if former queries made to all sellers associated with thresholds lower than $q + c_i$ have yielded prices greater than $q + c_i$. The seller using the Mean-Fonders Pricing heuristic will therefore set a price q^* according to:

$$q^* = argmax_q\big(q \prod_{r'_i \leq q+c_i}(1-F_i(q+c_i))\big) \quad (3)$$

3) Cost-Probability Tradeoff Pricing (M3): The intuition for using this Pricing method is the tendency of some people to find the balance between the probability of finding the commodity cheaper than the minimum commodity price found so far and between the price they are wiling to pay for this probability (i.e., is it worth paying $5 for finding the product in a different store with a 90% chance of being cheaper than the minimum price found so far).

The intuition for considering this exploration strategy relies on former evidence of people's use of thresholds (even in the form of rules of thumb) in their search behavior [23], [18], and at the same time experimental evidence indicating that people's exploration behavior does not seem to be related to risk aversion [27] but rather to loss aversion [24].

The Cost-Probability Tradeoff Pricing heuristic also assumes that the buyer follows a threshold-like decision rule; however, thresholds are dynamically set according to the best value obtained during any stage of the exploration. Specifically, the buyer is assumed to be calculating a threshold for each seller s_i, denoted r''_i, in the form of a weighted sum of the probability that the price that will be obtained from that seller is below the best price known, and the relative magnitude of the cost of querying that seller is comparable to the maximum possible querying cost:

$$r''_i(x) = (1-\alpha)P(q_i < x) + \alpha\frac{MaxCost - c_i}{MaxCost} \quad (4)$$

Where x is the minimal price obtained so far along the search, $MaxCost$ is the maximum possible querying cost and α is the weighting coefficient. At any stage of its exploration, the buyer will query the seller associated with the maximum threshold as long as it is greater than α which is, in fact, the threshold assigned to the seller associated with the minimum price among those obtained so far, according to (4).

Following the solution concept of the other two methods, the benefit-maximizing price q^* when buyers use the cost-probability tradeoff approach is given by:

$$q^* = argmax_q(q \prod_{r''_i(q+c)\geq\alpha}(1-F_i(q+c))) \quad (5)$$

Since the essence of the paper is not the pricing methods themselves, but rather the variation in the system performance when alternating between them, we do not get into a greater level of detail in their description. More details about the three price-setting methods can be found in the extended version of this paper, downloadable from: http://xrl.us/bmpouc.

B. Measures

The performance measure of interest when evaluating an exploration setting of the type used in this paper depends on the evaluator's goal. For example, if a seller's best interests are concerned, the seller's expected revenue should be the measure of interest. If a buyer's welfare is concerned then the buyer's exploration-related measures, such as the extent of search and expected overall expense, should be considered.

We use three different performance measures of the individual and system performance. These are the *exploration extent*, measuring the number of sellers a buyer explored throughout its exploration process; the buyer's *overall expense*, measuring the minimum value obtained plus the exploration costs accumulated along the process; and the *seller's revenue*, measuring the payment received by the seller using the evaluated pricing mechanism. If PDAs can exhibit behaviors similar to those of people in this domain, then all three measures should reflect a consistent pattern.

C. Experimental Infrastructure

Two experimental infrastructures were developed, simulating a price-search environment. The first was designed to experiment with PDAs and theoretic-optimal buyer and the second with people.

1) Evaluating PDAs: The PDAs' evaluation infrastructure enables our system to instantiate agents with the appropriate exploration problem input, receive their exploration choices and supply them with values based on their selections (according to the distribution of values of the different options available). The agents, upon receiving the problem input, which includes the costs of querying and the distribution of values, have to decide at each stage of the process whether to terminate the exploration, ending up with the lowest value revealed so far, or to resume exploration. In the case of continuing the exploration, they also have to inform the system who to query next.

To facilitate the evaluation of the pricing mechanisms with the theoretic-optimal buyer, an additional agent was developed applying the optimal exploration principles (i.e., using a set of thresholds according to Equation 1). This agent implemented the same API as the PDAs and thus could be used with the same experimental infrastructure.

Our experimentation used agents designed by computer science students in a core Operating Systems course. As part of her regular course assignment, each student programmed an agent according to the above guidelines. Each student's grade in the assignment was correlated with her agent's performance, i.e., the overall querying costs plus the value she ended up with. As part of their assignment, students provided documentation that described the algorithm used for managing the search. An external proxy program was used to facilitate communication with the different stores. The main functionality of the proxy was to randomly draw a store's price based on its distribution, if queried, and to calculate the overall querying costs and the price paid. The above procedure complies with the common practice in PDA-based research [4], [16], [15]. Overall, we used 31 PDAs that the students developed each tested with all of the problems from the set of problems described below with each of the three pricing methods.

2) Evaluating People: The second evaluation infrastructure developed is a JavaScript web-based application, emulating an exploration problem with 8 stores, each associated with a different distribution of prices and a cost for obtaining the true price (represented as a "parking cost" for parking next to that store). Figure 1 presents a screenshot of the system. In this example, the price of each of the two stores is known. Querying a store is done by clicking the "Check" button below it, in which case the true price of the store becomes known and the parking cost of that store is added to the accumulated cost. The game terminates when clicking the "Buy" button (available only in stores whose prices are known), upon which a short summary of the overall expense is presented to the player (divided into the accumulated exploration cost and the price paid for the product itself).

Subjects were recruited using Amazon's Mechanical Turk service[2]. When logged in, subjects received a short textual description of the experiment, emphasizing the exploration aspects of the process and the way costs are accumulated, followed by a short video clip. Then, a series of practice games were played in order to make sure that the subject understood the experiment. Participants had to play at least three practice games; however, they could continue practicing until they felt ready to start the experiment. Once finished practicing, the system randomly drew 10 problems from its problem repository of 100 problems and these were played sequentially. To prevent the carryover effect, a "between subjects" design was used, allowing each participant to participate only in one experiment using the same pricing method in all of the 10 problems presented to her. To motivate players to exhibit efficient exploration behavior, and possibly also extend their practice section, they were told that the 40% of them whom end up with the minimal average overall expense would receive a double payment for their participation in the experiment. As a means of precaution, the time it took each participant to make each selection and the overall time of each game played was logged, and participants with unusually low times were removed from the database. Overall, we had 150 people participating in our experiments, 50 for each pricing method (mechanisim).

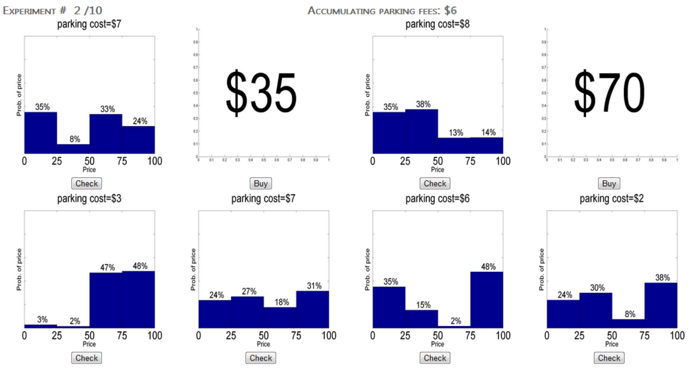


Figure 1: A screenshot of the system designed to experiment with people.

We note that PDA programmers and the Amazon Turk participants are not from the same population. The former are all computer science students. The latter are from a more general population. This choice of participants is intentional and complies with the general idea which motivates PDA literature, according to which the PDAs (whose development requires some knowledge in programming) can replace people (from the general population) in system evaluation. Therefore, our group of computer science students fairly represents the typical population of PDA programmers, whereas the group from Mechanical Turk corresponds to a general group of individuals one is likely to recruit for an experimental evaluation of a tested system.

D. Problem Sets

To simplify the search problem representation, distributions were formed as multi-rectangular distribution functions. In a multi-rectangular distribution function, the interval is divided into subintervals $x_0, .., x_n$, and the probability distribution is given by $f(x) = \frac{p_i}{x_i - x_{i-1}}$ for $x_{i-1} < x < x_i$ and $f(x) = 0$ otherwise, $(\sum_{i=1}^{n} P_i = 1)$. The benefit of using a multi-rectangular distribution function is its simplicity and modularity, in the sense that any distribution function can be modeled through it with a relatively small number of rectangles. Furthermore, the multi-rectangular distribution function is easier for people to grasp, as they can be given the explicit probability captured by each rectangle (equally distributed over the interval), as illustrated in Figure 1, rather than having them struggle with a distribution function which values are difficult to interpret.

The set of problems used consisted of 100 randomly generated problems. Each problem contained 8 stores, where one of them is the store for which the evaluated pricing techniques apply. The distribution of prices of the other seven stores were randomly generated under the constraint of having exactly four rectangles, each defined over an equal-size interval. This, again, in an effort to facilitate people's understanding of the price-distribution and their ability to reason about the problem. The overall interval of prices was $(0 - 100)$, as illustrated in Figure 1. The search cost for each store (including the store to which the pricing techniques are

[2] For a comparison between AMT and other recruitment methods see [20].

applied) was randomly picked from the interval $(1-10)$. Three variants of each problem were generated, one according to each of the pricing mechanisms described above.[3] The variants of each given problem thus differed only in the price set to the store for which the pricing applies.

V. Results

We briefly review the set of PDAs received and then describe the performance of PDAs compared to people as reflected in the experiments according to the different measures and evaluated mechanisms. Then we present a drill down analysis of per-problem individual performance obtained when using people and PDAs. Statistical significance was tested, whenever applicable, using a t-test, assuming unequal variance.

1) PDA strategy: The strategies used by the PDAs in our experiments reveal several characteristics along which agent designs varied, among which: relying on expected value, variance and the median price of each store, weighing the parking costs (querying costs) as an affecting factor, a preliminary selection of stores for querying, the inclusion of the cost incurred so far (i.e., "sunk cost") in the decision-making process and the use of the probability of finding a store with a lower price than the minimum found so far. It is notable that most of these characteristics do not affect the optimal strategy (see Section III). In particular, several of the PDAs (67%) use the mean price of a store as a parameter that directly influences the search strategy, even though the optimal strategy is not directly affected by means.

The analysis of people's strategies is beyond the scope of the paper for several reasons. First, while asking people to document or express their search strategy as part of the experiment is possible, relying on these statements is problematic, mostly because of the inability of people to reliably describe their strategies. Evidence of discrepancies between actual and reported human behavior has been widely cited throughout the paper. Second, reverse engineering of the collected data into a strategy is infeasible in our case due to the richness of the problems used and the limited number of observations collected for each person, as well as the large amount of possible different behaviors.

2) Performance Comparison: Figures 2-4 depict the average overall buyer's expense (payment for the product plus accumulated costs along the exploration), average exploration extent and seller's revenue, according to the type of buyer used (theoretic-optimal, person or PDA) for each of the different pricing mechanisims (M1, M2 and M3). The following table summarizes the similarity in terms of the average behavior between PDAs and people for each measure-mechanism combination:[4]

	M1	M2	M3
Overall expense	+ (0.34)	- (<0.001)	+ (0.065)
Exploration extent	- (<0.001)	- (<0.001)	- (<0.001)
Seller's revenue	- (<0.001)	- (<0.001)	+ (0.019)

As observed from the table, besides the case of using mechanism M2, there is no consistency in the determination of whether or not the performance of people is similar to the performance of PDAs across measures and across the tested mechanisms. For example, PDAs exhibit a similar average exploration expense and seller's revenue when the seller sets her prices according to M3; however, substantial differences are observed in the performance of the two groups in both measures when using M2. Similarly, there is a substantial difference in the seller's revenue when facing PDAs and when facing people, if using pricing mechanism M1 and M2; however, no significant difference is noted if using pricing mechanism M3.

As for comparing PDAs and people to the theoretic optimal agent, the performance of the latter was found to be substantially different $(p < 0.001)$ from those of PDAs and people in all measure-mechanism combinations. This result aligns with former literature [11], [18].

The substantial differences that were obtained for some of the measure-mechanism combinations when using PDAs and when using people suggest that individuals from the two populations use different exploration methods. To support this latter claim we present Figure 5, which supplies a drill-down comparison of the buyer's average expense according to their type and the individual problem used, for each of the price setting mechanisms. Notice that under each mechanism (different pricing method) the buyer faces a different problem, as the price of the seller whose price is set by the specific pricing method is different. For exposition purposes, the problems in each of the graphs are sorted according to the average performance achieved by human buyers, in ascending order. Each data point relating to people is the average of the overall expense achieved by human subjects who encountered the specific problem variant to which it relates (on the horizontal axis). The data points relating to PDAs depict the average performance of the 31 PDAs when encountering the same problem, and those relating to the theoretic-optimal agent depict the expected performance of the latter when given the problem. As can be observed from the figure, the average per-problem buyer's expense of PDAs is substantially different from the one obtained with people as buyers. Indeed, for some of the problems, the performance of PDAs is relatively close to that of human buyers (especially under M3); however, different results were obtained for the majority of the cases. As expected, the results of the theoretic-optimal agent are substantially better than the other two for the majority of the exploration problems tested. In a small portion of problems, people and/or PDAs managed to perform slightly better than

[3] For the M3, the value of α was arbitrarily set to 0.5.

[4] Similarity is determined whenever there is no statistically significant difference ($p < 0.001$) and is marked as "+". The number in brackets is the appropriate $p - value$.

the theoretic-optimal agent, this is attributed to the fact that the true price of each seller was drawn and fixed in each problem prior to running the experiment. If price is re-drawn from the distribution on each run, then the theoretic-optimal agent's average performance is always better when tested with a large set of problems. Similar patterns were obtained in the drill-down analysis according to problem type for the exploration extent and the seller's revenue measures.

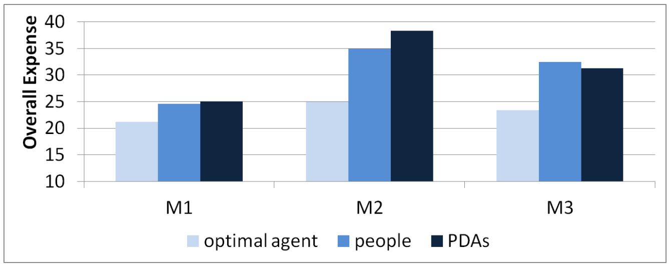


Figure 2: Average buyer's overall expense under the three pricing methods.

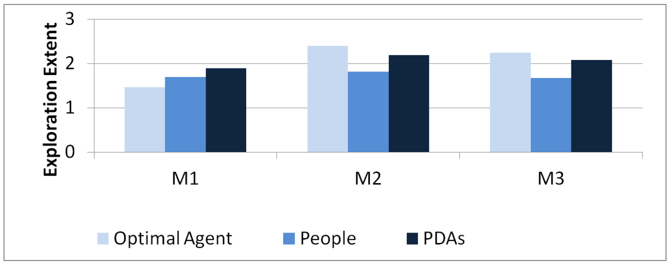


Figure 3: Average buyer's exploration extent under the three pricing methods.

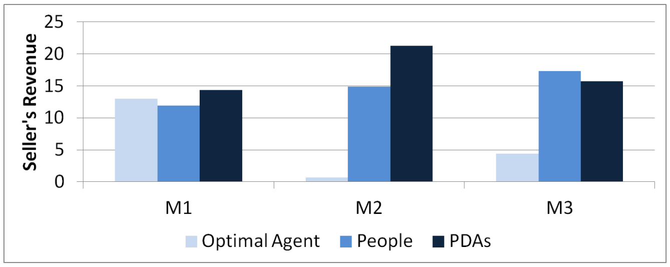


Figure 4: Average seller's revenue under the three pricing methods.

VI. Discussion and Conclusions

The results suggest that indeed the determination of whether or not PDAs can be used as a substitute for people for mechanism evaluation is mechanism and measure dependent. These findings are very different from claims made in prior work regarding the usefulness of the PDA-based approach and results reported concerning the similarity of people and PDAs in specific domains. Our results demonstrate the risks in generalizing based on the "average behavior" observed by applying a specific mechanism and using a specific set of measures for comparison. Thus, the use of PDAs should be carefully handled, and the similarity between the behavior of people and the PDAs used should be verified for every new mechanism that needs to be evaluated using the exact same measures of interest.

Specifically, in our case, had the seller tested the three methods with PDAs, the preferred pricing method would be Mean-Fonders Pricing (M2); whereas, if the buyers are people then the preferred pricing method should have been Cost-Probability Tradeoff (M3). Moreover, even if the seller decides to run a pilot test with one of the methods, evaluating the usefulness of using PDAs as a substitute for people in order to solve its price-setting problem, she could have reached the wrong conclusions. For example, if choosing the Cost-Probability Tradeoff (M3) pricing mechanism for the pilot study and using the expected revenue measure, the seller would assume PDAs can be used also for the evaluation of the other two methods and end up with the wrong choice. Even if the seller is given data on the extent of exploration and the average overall expense of PDAs in comparison to people (which are the two main measures that capture exploration), it should not assume that the use of PDAs with the same mechanisms would yield an expected revenue similar to the one obtained with human buyers.

Another interesting finding is that from the seller's point of view the dominant pricing method among the three methods evaluated, as reflected in Figure 4 is buyer-type-dependent (the Cost-Probability Tradeoff (M3) method is preferred when facing human buyers, the Mean-Fonders Pricing (M2) is preferred when facing PDAs and the Theoretic-Optimal Pricing (M1) is preferred when facing theoretic-optimal buyers). While this is not the essence of this paper, this result is important due to the nature of eCommerce — since sellers are likely to face a mixture of human and agent buyers, if the buyer type can be identified (e.g., using Captchas) a finer grained selection of the price setting mechanism can be used.

Finally, we note that despite the inability of generalizing the applicability of PDAs for mechanism evaluation, this technology is extremely useful in settings where it is found suitable as a substitute for people. Therefore, an important direction for future research is the development of methodologies for facilitating such evaluations.

References

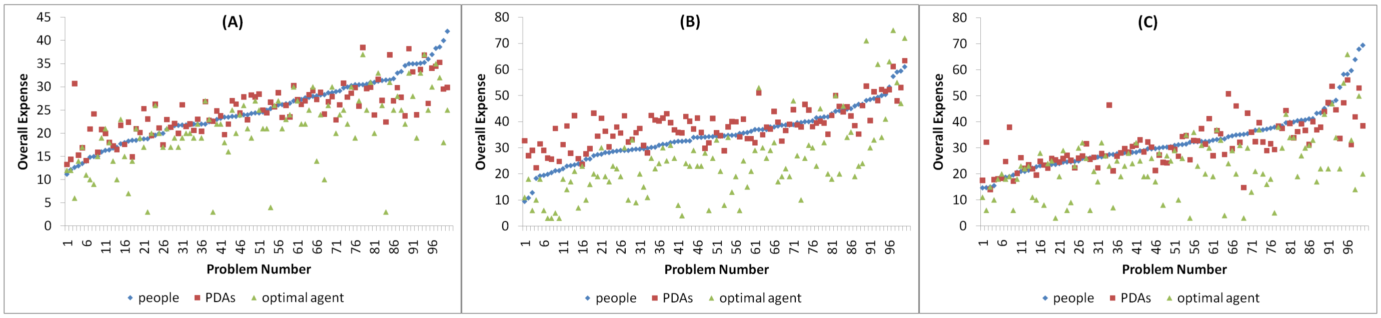
[1] M. Baye, J. Morgan, P. Scholten, and D. Jansen, "Persistent price dispersion in online markets," *The new economy and beyond: past, present and future*, 2006.

[2] M. Bertrand and S. Mullainathan, "Do people mean what they say? implications for subjective survey data," *American Economic Review*, vol. 91, no. 2, pp. 67–72, 2001.

[3] M. Chalamish, D. Sarne, and S. Kraus, "Mass programmed agents for simulating human strategies in large scale systems," in *AAMAS*, 2007, pp. 135–136.

[4] ——, "Programming agents as a means of capturing self strategy," in *AAMAS*, 2008, pp. 1161–1168.

[5] A. Chavez and P. Maes, "Kasbah: An agent marketplace for buying and selling goods," in *PAAMS96*, 1996, pp. 75–90.



(a) M1 (b) M2 (c) M3

Figure 5: The average expense of buyers over the 3 pricing methods: (a) M1; (b) M2; and (c) M3.

[6] K. Clay, R. Krishnan, E. Wolff, and D. Fernandes, "Retail strategies on the web: Price and non-price competition in the online book industry," *J. of Ind. Economics*, pp. 351–367, 02.

[7] S. Gal, M. Landsberger, and B. Levykson, "A compound strategy for search in the labor market," *International Economic Review*, vol. 22(3), pp. 597–608, 1981.

[8] B. Grosz, S. Kraus, S. Talman, B. Stossel, and M. Havlin, "The influence of social dependencies on decision-making: Initial investigations with a new game," in *AAMAS*, 04, pp. 780–787.

[9] C. Harries, J. Evans, and I. Dennis, "Measuring doctors' self-insight into their treatment decisions," *Applied Cognitive Psychology*, vol. 14, pp. 455–477, 2000.

[10] N. Hazon, Y. Aumann, and S. Kraus, "Collaborative multi agent physical search with probabilistic knowledge," *IJCAI*, pp. 167–174, 2009.

[11] J. Hey, "Still searching," *Journal of Economic Behavior and Organization*, vol. 8, no. 1, pp. 137 – 144, 1987.

[12] J. Jumadinova and P. Dasgupta, "Firefly-inspired synchronization for improved dynamic pricing in online markets," *SASO*, pp. 403–412, 2008.

[13] G. Kempen, M. van Heuvelen, R. van den Brink, A. Kooijman, M. Klein, P. Houx, and J. Ormel, "Factors affecting contrasting results between self-reported and performance-based levels of physical limitations," *Age and Ageing*, vol. 25, no. 6, pp. 458–464, 96.

[14] J. Kephart, J. Hanson, and A. Greenwald, "Dynamic pricing by software agents," *Computer Networks*, vol. 32, no. 6, pp. 731 – 752, 2000.

[15] R. Lin, S. Kraus, N. Agmon, S. Barrett, and P. Stone, "Comparing agents: Success against people in security domains," in *AAAI*, 2011, pp. 809–814.

[16] R. Lin, S. Kraus, Y. Oshrat, and Y. Gal, "Facilitating the evaluation of automated negotiators using peer designed agents," in *AAAI*, 2010, pp. 817–822.

[17] P. Maes, *Designing Autonomous Agents*. Cambridge, MA: MIT Press, 1990, pp. 105–122.

[18] P. Moon and A. Martin, "Better heuristics for economic search, experimental and simulation evidence," *Behavioral Decision Making*, vol. 3, no. 3, pp. 175–193, 1990.

[19] T. Offerman, J. Potters, and H. Verbon, "Cooperation in an overlapping generations experiment," *Games and Economic Behavior*, vol. 36, no. 2, pp. 264–275, 2001.

[20] Paolacci, Gabriele, Chandler, Jesse, and Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[21] A. Rosenfeld and S. Kraus, "Modeling agents based on aspiration adaptation theory," *Autonomous Agents and Multi-Agent Systems*, vol. 24, no. 2, pp. 221–254, 2012.

[22] D. Sarne, A. Elmalech, B. Grosz, and M. Geva, "Less is more: Restructuring decisions to improve agent search," in *AAMAS*, 2011, pp. 431–438.

[23] A. Schotter and Y. Braunstein, "Economic search: an experimental study," *Economic Inquiry*, vol. 19, no. 1, pp. 1–25, 1981.

[24] D. Schunk and J. Winter, "The relationship between risk attitudes and heuristics in search tasks: A laboratory experiment," *Economic Behavior Organization*, pp. 347 – 360, 09.

[25] R. Selten, K. Abbink, J. Buchta, and A. Sadrieh, "How to play (3 x 3)-games.: A strategy method experiment," *Games and Economic Behavior*, vol. 45, no. 1, pp. 19–37, October 2003.

[26] R. Selten, M. Mitzkewitz, and G. Uhlich, "Duopoly strategies programmed by experienced players," *Econometrica*, vol. 65, no. 3, pp. 517–555, 1997.

[27] J. Sonnemans, "Strategies of search," *Journal of Econ. Behavior and Organization*, vol. 35, no. 3, pp. 309 – 332, 1998.

[28] Y. Wan and G. Peng, "What's next for shopbots?" *IEEE Computer*, vol. 43, pp. 20–26, 2010.

[29] M. Weitzman, "Optimal search for the best alternative," *Econometrica*, vol. 47, no. 3, pp. 641–54, May 1979.

[30] M. P. Wellman, P. R. Wurman, K. O'Malley, R. Bangera, S. Lin, D. Reeves, and W. E. Walsh, "Designing the market game for a trading agent competition," *IEEE Internet Computing*, vol. 5, no. 2, pp. 43–51, 2001.

[31] Y. Zhang, K. Biggers, L. He, S. Reddy, D. Sepulvado, J. Yen, and T. Ioerger, "A distributed intelligent agent architecture for simulating aggregate-level behavior and interactions on the battlefield," in *SCI*, 2001, pp. 58–63.