Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Leveraging Fee-Based, Imperfect Advisors in Human-Agent Games of Trust

Cody Buntain
Department of Computer Science
University of Maryland
College Park, Maryland 20742 USA

Amos Azaria and **Sarit Kraus**
Department of Computer Science
Bar-Ilan University, Ramat-Gan, Israel 52900

Abstract

This paper explores whether the addition of costly, imperfect, and exploitable advisors to Berg's investment game enhances or detracts from investor performance in both one-shot and multi-round interactions. We then leverage our findings to develop an automated investor agent that performs as well as or better than humans in these games. To gather this data, we extended Berg's game and conducted a series of experiments using Amazon's Mechanical Turk to determine how humans behave in these potentially adversarial conditions. Our results indicate that, in games of short duration, advisors do not stimulate positive behavior and are not useful in providing actionable advice. In long-term interactions, however, advisors do stimulate positive behavior with significantly increased investments and returns. By modeling human behavior across several hundred participants, we were then able to develop agent strategies that maximized return on investment and performed as well as or significantly better than humans. In one-shot games, we identified an ideal investment value that, on average, resulted in positive returns as long as advisor exploitation was not allowed. For the multi-round games, our agents relied on the corrective presence of advisors to stimulate positive returns on maximum investment.

Introduction

As humans, we often rely on the advice of family, friends, or supposed experts when making decisions where we have incomplete and imperfect information. Such interactions require some level of trust in both the advisor and the service provider despite threats of exploitation and collusion. Regardless of these risks, people tend to trust others even when they have no solid justification. Existing research even shows individuals are often rewarded for their trust with reciprocation (Berg, Dickhaut, and McCabe 1995). In the digital realm, however, agents acting on our behalf may not be able to rely on reciprocity and must carefully weigh decisions on who or what to trust. Fortunately, our increasingly connected and digitized world allows agents access to a vast array of information sources to support such decisions. As in the real world, such resources are not immune to compromise or manipulation: a malicious seller might employ bots or pay people to drive up his reputation on a review site, and when a buyer makes a particularly expensive purchase with this seller, the seller takes the money and disappears. Similarly, malicious parties may post advice in forums on upcoming spikes in stock or currency values to support pump-and-dump schemes to trick people into participating in an otherwise worthless investment.

While a large body of work has focused on building better recommendations systems to address these and related issues (Jøsang, Ismail, and Boyd 2007), we instead explore how (or whether) a buyer, human or agent, could leverage these resources given their potentially adversarial or compromised nature. Rather than focus on building better information sources, we investigate how these imperfect sources can be integrated into an agent's decision-making process. Further complicating matters, reliance on these external sources might be costly (time, money, or resources), so an agent cannot simply query all possible information sources before making a decision. Trusting third-party input is then a tradeoff between the potential gain in utility from better information and the loss incurred from cost, noise, and potentially low integrity. To explore this balance, we developed a set of game constructs in which these potentially costly and imperfect third parties can be evaluated.

By extending Berg's investment game to include costly and imperfect/exploitable advisors, we can model human interactions with these advisors, determine whether including advisors facilitates or hinders trust reciprocity, and determine how an agent might best use these resources. Coupling this construct with Amazon's Mechanical Turk[1] then allows us to collect data across hundreds of interactions with human participants to gain insight into human behaviors. We then use this behavioral data to estimate parameters like average

[1] www.mturk.com

returns and exploitation likelihoods to determine investment amounts and advisor solicitation strategies that maximize returns and compare our agents against human performance.

Related Work

In the mid-nineties, Joyce Berg and her colleagues introduced a construct to study trust and reciprocity with a simple two-player investment game (Berg, Dickhaut, and McCabe 1995). This investment game has been a popular lens through which researchers have explored methods to support or suppress trust between players. Berg brought together a number of undergraduate students to play a double-blind, one-shot game in which players were divided into two groups: investors and investees (Berg et al. used "trustors" and "trustees"). All individuals were given an initial sum of $10 for participation. Investors then chose some amount of this $10 sum to send to, or "invest with", the investee. The investee then received an amount triple the investment sent by the investor and decided how much of this tripled amount to return to the investor. Investees were said to reciprocate trust if the amount returned was greater than the investor's investment. Results showed a non-zero average investment and return, which ran counter to the subgame perfect equilibria in which the investee returns nothing, and the investor, expecting the investee's behavior, invests nothing.

In 2004, Cochard et al. investigated an extension to Berg's game in which players were allowed to play multiple rounds and comparing a seven-round repeated game in which the number of rounds were known a prior with Berg's one-shot version (Cochard, Van, and Willinger 2004). Results from this experiment indicated that investors invest significantly larger portions of their initial sums and investees pay back significantly larger portions of the multiplied investments in the first five rounds. In the last two rounds, however, investee reciprocity dropped sharply, which resulted in a reduction in investor investment (consistent with Anderhub's earlier findings (Anderhub, Engelmann, and Güth 2002)). Cochard theorized investor behavior was consistent with the "reciprocity hypothesis" throughout the game whereas investee behavior conformed only in early stages before reverting to strategic, self-interested play. At the same time, Engle-Warnick and Slonim were also investigating multi-round modifications to Berg's game by exploring the differences between investment games where the number of rounds were known a priori against games that could continue indefinitely (Engle-Warnick and Slonim 2004). These possibly infinite games had a probability p of continuing or $1-p$ of terminating after each round. Engle-Warnick and Slonim concluded that trust and reciprocity decrease with each round when the number of rounds were known but remained relatively constant if the number of rounds could vary. Additionally, they observed that investor trust resets to the player's default level (governed by an external propensity to trust) when paired with a new partner.

More related to our research, however, are the works of Bracht and Feltovich, which allowed cheap talk between investors and investees and the investor observation of previous investee behavior (Bracht and Feltovich 2009). They concluded that, while cheap talk had little effect, observation of the investee's behavior in the previous round "improves aggregate cooperation and efficiency substantially, and significantly, over the basic trust game." Bracht extended this work in which observations of prior investee behavior were provided by a human third-party who was not guaranteed to provide accurate information (Bracht 2010). Despite potential inaccuracy, Bracht found this third-party observer still stimulated trust. Though this result is interesting and relevant, our work departs from Bracht and Feltovich's by attaching a cost to observer information and allowing investees to bribe the observer for a fee.

This paper also touches on research into agents in games of trust and cooperation. Hendrix et al. studied how reputation information influenced agent behavior in cooperative games (Hendrix and Grosz 2007). While Hendrix's experiments did not include actual human behavior, he did account for costly and manipulated information similar to us; as one would expect, he found agents would forego soliciting or integrating reputation as information costs increased and the likelihood of incorrect reputation information increased. His use of reputation is similar to how our advisors generate advice for investors, so we should also consider related work on reputation systems, on which Jøsang et al. provide a good survey (Jøsang, Ismail, and Boyd 2007). Jøsang's work considers many different types of reputation systems in online environments and provides a listing of deployed uses of this technology. More interestingly, Jøsang includes details on how reputation systems stimulate a collaborative sanctioning effect that incentivizes positive behavior. He also discusses how such systems might be compromised by malevolent actors. Though our advisor construct shares similarities with reputation systems, research into reputation systems seeks to increase quality of the information provided whereas we address issues of quality from a different perspective. We instead assume the existence of these potentially high-quality systems as resources external to our agents and account for the boundless human ability to co-opt such systems, a point of view supported by Salehi-Abari's recent work on con-man agents that explicitly target and exploit trust and reputation models (Salehi-Abari and White 2010).

Finally, parallels exists between the work described herein and the work of Azaria et al. on automated advice provisioning in repeated human-agent interactions (Azaria et al. 2012). Both explorations investigate effects of automated advice on human-agent interactions, but whereas Azaria's advice-giving agent seeks to influence the human to select an action that satisfies the utilities of both the human and agent, our work instead seeks to identify strategies that can maximize human/investor utility given self-interested or compromised advisors.

Extended Game Construct

Before describing our experiments, we first present the extensions we made to Berg's investment game, which we refer to as the "Advisor Game." As the name suggests, the primary difference is the addition of k automated advisor resources that provide advice to investors. In each round, these advisors observe the difference between the investor's

investment M_I and the amount the investee returns to the investor M_R. When an advisor is solicited for advice, it will recommend investment if this difference is non-negative ($M_I - M_R \geq 0$) or not to invest if this difference is negative ($M_I - M_R < 0$). Advisor observations are noisy such that their advice will be incorrect with probability P_n (we included noise to avoid complete information revelation when an advisor advises incorrectly or two advisors disagree).

Prior to selecting an amount to invest from the investor's initial wealth W, the investor can choose to solicit advice from any number s of these k advisors (selected advisors are chosen at random). Furthermore, when the investor solicits advice in a given round, he pays a solicitation fee ρ_s to each advisor solicited (this fee is a percentage of the investor's total for that round). For example, if the solicitation fee is 10% and the investor solicits advice from 3 investors, invests all \$10, and receives \$20 from the investee, he must pay \$2 to each of the 3 investors. Before the investor solicits advice, however, the investee can choose to bribe any number b of these advisors for a fee ρ_b (this bribery fee is also a percentage of the investee's total at the end of the round). When an advisor is bribed, that advisor will always recommend investment regardless of noise. The investee has no control over which advisors he bribes, so if he bribes two advisors and the investor solicits advice from two advisors, there is no guarantee the two bribed advisors will be solicited for advice. These fees are calculated and removed from the players' totals at the end of each round.

Remaining game parameters follows those of Berg and Cochard. For instance, after bribery and solicitation, the investor invests M_I, the investee receives three times that investment $3M_I$, and he returns some amount M_R to the investor such that $0 \leq M_R \leq 3M_I$. We also provide the investor with an initial sum of $W = \$10$ at the start of every round, and his winnings at the end of each round are set aside and are not used as part of the investment in the next round.

Empirical Methodology

Our goals with this paper were to answer whether costly and imperfect advisors contribute to or hinder trust reciprocity and to determine how an automated agent might leverage these resources. To that end, the experiments described covered several conditions to compare player behavior across one-shot and multi-round games, each with three types of games: without advisors, with advisors but no bribery, and with advisors plus bribery. To provide advisors with the information necessary to make observations, each experiment included a priming phase and testing phase. The priming phase employed Berg's original one-shot investment game and paired the human player with an automated agent that followed the aggregate behavior from Berg's paper. If the human player was an investor, this agent returned a value calculated from the linear model $\max\{0, 2.1 \cdot I - 7\}$, which returns a non-negative return if the investor invested more than 6.36, as derived from data in Berg's experiment. If the human player was an investee, the agent invested \$7, chosen because the majority of investees in Berg's game provided positive returns given such an investment. Advisors then used the investee's return behavior to prime their observations for the testing phase. It is worth noting that human players were not told this first phase would be with an automated agent.

Following priming, testing-phase games were selected from one of the three advisor game types. During the initial data-gathering experiments, these advisor games were played between human players with the exception of timeouts. Timeouts occurred when players left the experiment early or in situations where a second human player could not be found. To gather data in these circumstances, players were paired with an automated player with a purely random strategy. Players who played with a majority of these random agents were excluded from analysis. Remaining analysis used a two-tailed Mann-Whitney-Wilcoxon (MWW) U test with $\alpha < 0.05$.

Game Conditions

As mentioned, one-shot and multi-round experiments included three game types: no advisors, with advisors but no bribery, and with advisors and bribery. In all cases, the number of advisors was $k = 5$, advisor noise was $P_n = 0.01$, and bribery cost was $\rho_b = 0.1$. For one-shot games, solicitation cost was constant at $\rho_s = 0.1$, but we varied it between $\rho_s = 0.01$ and $\rho_s = 0.1$ in multi-round games. Also in one-shot games, players were told they would play one round in each phase with different players. Players in multi-round games were told the priming phase would have one round, and the testing phase would have between three and eight rounds, each of which would include a new partner. When matching new players, our system made no attempt to synchronize rounds (that is, a player in round 3 could be paired with a player in round 5), and players were never told how many rounds their partners had played. Prior to game play though, participants were given instructions on how their games would be played and were then required to pass a short quiz to demonstrate understanding of game rules.

Participant Selection

To obtain a large and diverse data set in a cost-effective manner, we leveraged Amazon's Mechanical Turk framework to recruit participants via the Web. These players were paid \$0.20 for participating in the experiment and were given a bonus proportional to their winnings as an additional incentive to play well (the bonus was 1% of the total dollar amount earned in the one-shot game and 1% of the average per-round total in the multi-round game).

Each human player accepted a Human Intelligence Task (HIT) from the Mechanical Turk interface in order to take part in this experiment, which allowed us to enforce a number of qualifications prior to play. First, players were only allowed to take part in the experiment once for the one-shot game and once for the multi-round game. Second, all players were required to have a high approval rating (greater than 96%) from other Mechanical Turk requesters in order to take part. Third, all players must have completed more than 60 HITs in Mechanical Turk. Lastly, players were required to be in the United States to minimize any cultural bias.

Mechanical Turk provided a large sample of more than 1000 players that decomposed into the following demographics: approximately 60% male and 40% female, primarily between the ages of 18-34 (33% of participants were 18-24 and 44% between 25-34, with the remainder between 35-74), and mostly split between high school and baccalaureate degrees (49% and 42% respectively with the remaining 9% having post-baccalaureate degrees). While Mechanical Turk allowed us to restrict the experiment to those users within the United States, we asked participants to disclose their country of birth to account for cultural bias, 94% of which were also born in the United States.

Modeling Investment Expectations

To aid in designing strategies for our agents, we developed a model to calculate the expected return on investment $E^{-}_{s,c}[u]$ for an investor who solicits s advisors, at least c of which recommend investment, with solicitation fee ρ_s and invests M_I as shown in Eq. 1. The $E_{s,c}[u]$ term expresses this expected return before solicitation fees are deducted.

$$E^{-}_{s,c}[u] = (1 - s\rho_s)E_{s,c}[u] \quad (1)$$

Before further deconstructing investor expectation, we introduce some notation: t denotes the current round, $t-1$ denotes the previous round, and investee behavior in round t is given by the indicator function $X_t = \mathbf{1}_t(M_I - M_R \geq 0)$. If $X_t = 1$, we say the investee behaved positively; otherwise, if $X_t = 0$, the investee is said to have behaved negatively. We also define ω as the number of advisors who recommend investment such that $P(\omega \geq c)$ denotes the probability that at least c advisors say to invest ($0 \leq c \leq s$). Finally, we denote the current round's average rates of return for investees who behaved positively in $t-1$ as γ_+ and negatively in $t-1$ as γ_-. The $E_s[u]$ function is then a convex combination of the three advisor cases an investor will encounter: advice against investing, correct advice to invest given a positive investee, and incorrect advice to invest given a negative investee (Eq. 2).

$$\begin{aligned} E_{s,c}[u] = {} & (1 - P(\omega \geq c))W \\ & + P(\omega \geq c, X_{t^-1}{=}1)\left((W - M_I) + \gamma_+ M_I\right) \\ & + P(\omega \geq c, X_{t-1}{=}0)\left((W - M_I) + \gamma_- M_I\right) \end{aligned} \quad (2)$$

When soliciting no advisors, the model loses its dependence on the previous round and simply reduces to the product of the investment M_I with the average rate of return in the current round $\hat{\gamma}$ plus the remaining initial sum: $E^{-}_{0,0}[u] = (W - M_I) + \hat{\gamma}M_I$.

Results and Analysis

To reiterate, these experiments sought to answer whether advisors were beneficial in games of trust and provide data for developing effective agent strategies. Before discussing results specific to the one-shot and multi-round experiments, we identified four results present across all of our experiments. First, investors leveraged on average one advisor across all rounds. Second, where possible, investees bribed on average approximately one advisor, even if doing so was unnecessary (that is, an investee might bribe an advisor regardless of his previous behavior). Thirdly, owing to investor and investee propensity to spend funds on advisors, the presence of advisors significantly increased social welfare (defined as the sum of dollars spent). Finally, when an advisor recommended investing in either the no-bribery or bribery scenarios, investors invested significantly more than investors without advisors and significantly more than when advisors recommended against investing. These results were statistically significant by the MWW test at $\alpha < 0.05$.

One-Shot Games

For one-shot games, we had 418 participants, and the presence of advisors had no significant effect on aggregate investee behavior in either the priming or the testing phases. In the no-advisor case, our experiments showed investees returned on average 6% below an investor's investment in the testing phase. Similarly, investees in the no-bribery and bribery cases returned on average 4% and 10% below the investment with p-values both higher than 0.05, showing no significant difference from the no-advisor case. In the no-bribe scenario, constrained analysis of those investees who behave positively in the priming phase showed the average return increased to 5% above the investment. An MWW test of this data, however, did not show a statistically significant difference between these returns and the returns of the no-advisor case ($p = 0.54$). When bribery was allowed, average returns from positively behaving investees decreased to 4% below the investment. If we then accounted for investees who bribed advisors, that number then further decreased to 11% below.

Furthermore, according to our data, limited correlation existed between positive investee behavior in the priming phase and positive behavior in the testing phases across all three game types (the Pearson correlation coefficients were $r = 0.44$, $r = 0.47$, and $r = 0.37$ for no-advisor, no-bribe, and bribe cases respectively). As such, advisor advice was not a good indicator of investee behavior in the current round. From these results, it follows that investors did not benefit from soliciting advisors for advice regardless of whether bribery was allowed. Investors who did solicit advice performed equally as well as investors without advisors but then paid solicitation fees on top of their equivalent returns, which resulted in significantly lower totals. These conclusions were consistent with Cochard's findings in that end-game dynamics seemed to dominate investee behavior.

Multi-Round Games

Unlike the one-shot games, data from our 147 participants indicated advisors stimulated significantly positive effects on both investor and investee behavior in the multi-round games regardless of bribery (bribe fee was fixed at $\rho_b = 0.1$ of investee winnings) and solicitation fee (tested at $\rho_s = 0.1$ and $\rho_s = 0.01$ of investor winnings). Of these effects, likely the most important was that investees returned significantly higher percentages of investment when advisors were present than when they were absent, regardless of bribery. These results are summarized in Table 1, which shows the mean returns with $\rho_s = 0.1$ and the p-value of the MWW

test against the no-advisor games. From this table, one can also see that investee behavior without advisors was consistent with investee behavior in the one-shot games.

Table 1: Investee Return Percentages with $\rho_s = 0.1$

Type	Players	Mean Return	p-Value
No Advisor	22	-4%	
No Bribery	31	15%	0.0349
Bribery	20	22%	0.0155

Investors then behaved consistently with Cochard's reciprocity hypothesis and invested significantly more in games with advisors than in games without. Table 2 illustrates this behavior with average investments for each game type and the p-value of the MWW test against the no-advisor games (also with a solicitation fee $\rho_s = 0.1$). One may see the number of investors and investees are not equal, which was an artifact of player timeouts.

Table 2: Mean Investment with $\rho_s = 0.1$

Type	Players	Mean Investment	p-Value
No Advisor	23	4.8	
No Bribery	30	5.8	0.0238
Bribery	21	6.6	0.0002

Despite these benefits, it is not clear whether our advisors provided worthwhile advice. As in the one-shot games, correlation between investee behavior in the previous round to the current round was still relatively low ($r = 0.4$ for no advisors, $r = 0.47$ for no bribery, and $r = 0.46$ for bribery). Instead, investees were more likely to behave positively in the current round regardless of their previous behavior, which seems consistent with the collaborative sanctioning effect of a reputation system. This effect is apparent in the large difference in investment between an investee who behaved negatively in the previous round and an investee who behaved positively when advisors were present, as shown in Figure 1 (the non-zero difference in the no-advisor case was unexpected and may have been an artifact of investors being re-paired with the same investees when few players were available). These large differences suggest advisors allow investors to punish non-reciprocating investees with smaller investments. This investor behavior was also present in the one-shot round but did not have the opportunity to affect investee behavior given the short interaction period.

We also explored multi-round games with solicitation fees that were much lower than bribery fees ($\rho_s = 0.01$ instead of $\rho_s = 0.1$). Aggregate behavior among investors and investees were unaffected by this reduction in cost, but the difference in investment between positive and negative investees was more pronounced.

Agent Strategies

In light of these results, our agent strategies differed between the one-shot and multi-round experiments. For the one-shot games, the low correlation between behavior in the

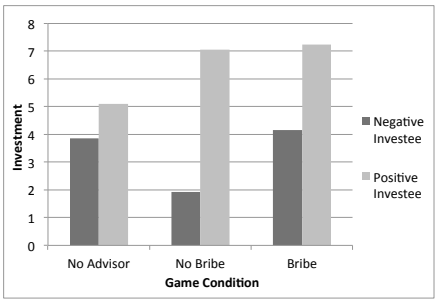

Figure 1: Average Investment Given Investee Behavior in the Previous Round

priming and testing phases dampened advisor efficacy, and our data demonstrated that investors who leveraged advisors performed below investors who did not solicit advisors. As such, a successful agent strategy should not solicit any advise in these one-shot games. This conclusion left then left only one question: how much should our agent invest?

Cochard's reciprocity hypothesis would suggest investing the largest possible amount, but our data demonstrated investees do not adhere to this hypothesis. A number of human investors also invested all of their initial sums and lost out on the return, which further suggested our agent should avoid investing its entire sum. Therefore, we used regression to model the relation between investment and return to find a value that maximizes our return; if such a value was negative, our agent should follow the sub-game perfect equilibrium and not invest at all. Since linear models would exhibit maximum values at investment extremums, we instead modeled the investment-return relationship with quadratic polynomials. These models showed a maximum positive return around an investment of \$6 for no-advisor and no-bribery games; for games with bribery, however, no investment result in a positive return. Agents based on these models performed equally as well as human players in the no-advisor game (means of 10.4 and 10.11 respectively with $p = 0.7949$) and performed significantly better than humans in the no-bribe game (means of 10.77 and 8.87 respectively with $p < 0.05$). In the bribery game, our agent would always end with 10, which is significantly higher than the average human player's payoff of 8.02. Figure 2 illustrates these results.

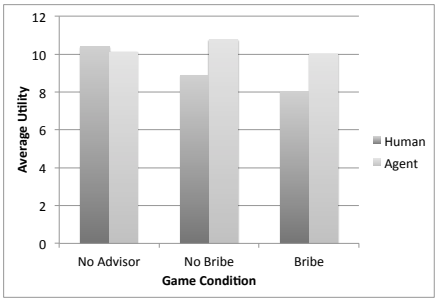

Figure 2: One-Shot Investor Utility (higher is better)

In the multi-round game, our agent was in a better position. While correlation between investee behavior in the previous and current rounds was still low, soliciting advice might still be worthwhile, so we turned to our model of expected investment return. First, data from games without bribery revealed that investees who behaved positively in round $t-1$ returned $\gamma_+ = 1.2$ or 120% of the investment, and negative investees returned $\gamma_- = 0.49$ or 49% of the investment. Then, by disregarding noise and calculating expected investment return following one advisor ($s = c = 1$), we calculated $P(\omega \geq c) = 0.78$ and established an upper bound on investor utility with advisors: if our agent solicited advice from a single advisor and invested $M_I = 10$, its expected return was 10.43. For no investment, our agent was guaranteed a utility of 10, and if our agent invested $M_I = 10$ and asked for no advice, it should expect a payoff of 10.45. Therefore, in the no-bribe case, the dominant strategy was to invest everything and ask for no advice.

Multi-round games with bribery were slightly more complex given the larger probability of faulty advice, which we accounted for by investing only when all s advisors recommended that action. As in the no-bribe games, data from our experiments revealed $\gamma_+ = 1.2$, $\gamma_- = 0.63$, and $P(X_{t-1} = 1) = 0.8$. Table 3 shows the empirical probabilities and expected returns pre- and post-fee given s agents recommending investment. As with the no-bribe game, one can see the dominant strategy for our investor agent is again to invest the entire sum without soliciting advice. It is also clear from this table that, while soliciting more advisors increases investor utility, the solicitation fee makes doing so cost prohibitive. To explore this line of inquiry further, the additional experiment with a reduced solicitation fee $\rho_s = 0.01$ resulted in the expectations shown in Table 4. Once again, soliciting no advisors provided the highest expected investor return, which may be the result of more severe punishment brought on by investors' soliciting more advisors.

Table 3: Investor Returns, $M_I = 10, \rho_s = 0.1, c = s$

s	$P(\omega \geq c, X_{t-1} = 0)$	$E_{s,c}[u]$	$E^-_{s,c}[u]$
0	–	11.18	11.18
1	0.07	11.36	10.23
2	0.05	11.45	9.16
3	0.04	11.51	8.06
4	0.03	11.53	6.92
5	0.02	11.56	5.78

Table 4: Investor Returns, $M_I = 10, \rho_s = 0.01, c = s$

s	$P(\omega \geq c, X_{t-1} = 0)$	$E_{s,c}[u]$	$E^-_{s,c}[u]$
0	–	11.24	11.24
1	0.05	11.25	11.13
2	0.03	11.37	11.14
3	0.01	11.45	11.10
4	0.004	11.50	11.03
5	0	11.53	10.95

Multi-Round Agent Strategy Results

To determine the efficacy of these agent strategies, we ran a series of experiments that matched these agents against human investees. To ensure these agents played in an environment consistent with the previous experiments, we also developed a human-like agent that behaved in a manner similar to human investors, so investees would be exposed to multiple strategies. This human analog solicited advice from a single advisor and if advised to invest, it would select a random investment from a triangular distribution with a minimum at $M_I = 5$ and maximum at $M_I = 10$. If the advisor advised against investment, this agent would then select a random investment from a triangular distribution with minimum at $M_I = 5$ and maximum at $M_I = 0$ (including this human analog agent also likely stimulated collective sanctioning seen in the human experiments). Along with these human analogs, our investor agents were able to outperform human investors regardless of whether bribery was allowed with an average return of 10.87 without bribery and average of 11.12 when bribery is allowed and solicitation fee is $\rho_s = 0.1$ (both with $p < 0.01$). Figure 3 demonstrates these superior agent results.

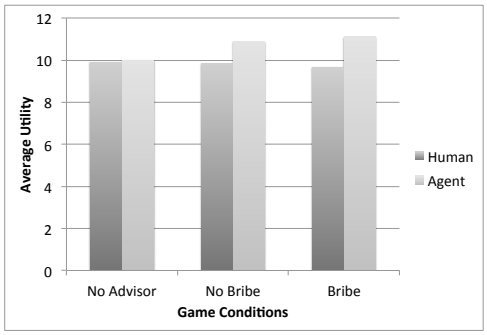

Figure 3: Multi-Round Investor Utility (higher is better)

Conclusions

We sought to answer whether the presence of imperfect advisors would enhance investor performance in games of trust, and if so, how an automated agent might leverage these additional resources to make better informed decisions. For games with limited interactions between investors and investees (as in our one-shot games), advisors do not seem to provide much benefit. With interactions of longer duration, however, the mere presence of advisors stimulates higher returns from investees, which in turn stimulates more investment from investors. This result is particularly interesting as it suggests collective sanctioning evolves even if the reputation system stimulating it is imperfect. Such interactions play out daily when people purchase goods and services via the Web. Based on our results, as long as sellers are monitored by advisors and some portion of the population follows their advice, we have shown agents can place trust in these sellers without needing to incur solicitation costs.

Extensions of this research could include advisors that provide more complex information about multiple investees

rather than the binary advice in these experiments. While this information would require additional cognitive cost to integrate, it might also complicate advisor manipulation. Additionally, one might leverage existing models of information gathering actions to calculate the maximum allowable cost for an advisor to be beneficial.

Acknowledgments

This work was supported in part by ERC grant #267523.

References

Anderhub, V.; Engelmann, D.; and Güth, W. 2002. An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization* 48(2):197–216.

Azaria, A.; Rabinovich, Z.; Kraus, S.; Goldman, C. V.; and Gal, Y. 2012. Strategic Advice Provision in Repeated Human-Agent Interactions. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, number Camerer 2003, 1522–1528.

Berg, J.; Dickhaut, J.; and McCabe, K. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10(1):122–142.

Bracht, J., and Feltovich, N. 2009. Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game. *Journal of Public Economics* 93(910):1036–1044.

Bracht, J. 2010. Trusting Your Sources. *The economist* 4(4):0–8.

Cochard, F.; Van, P. N.; and Willinger, M. 2004. Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization* 55(1):31–44.

Engle-Warnick, J., and Slonim, R. L. 2004. The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization* 55(4):553–573.

Hendrix, P., and Grosz, B. J. 2007. Reputation in the Venture Games. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, 1866–1867. AAAI Press.

Jøsang, A.; Ismail, R.; and Boyd, C. 2007. A survey of trust and reputation systems for online service provision. *Decision support systems* 43(2):618–644.

Salehi-Abari, A., and White, T. 2010. Trust Models and Con-Man Agents: From Mathematical to Empirical Analysis. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* 842–847.