Trading Agents Competing: Performance, Progress, and Market Effectiveness

Michael P. Wellman and Shih-Fen Cheng

University of Michigan
Artificial Intelligence Laboratory
Ann Arbor, MI 48109-2110 USA
{wellman,chengsf}@umich.edu

Abstract

How good are TAC-02 agents? How effective is the TAC travel market?

1 Introduction

One of the primary motivations for producing the original Trading Agent Competition (TAC) was to spur research on a common problem, thus enabling researchers to compare techniques and build on each others' ideas [Wellman and Wurman, 1999]. Working on a shared problem coordinates attention on particular issues (among the many of interest in the trading domain), and facilitates communication of methods and results by fixing a set of assumptions and other environment settings. An inspiring example for the TAC developers was the 1990 Santa Fe Double Auction Tournament [Rust *et al.*, 1993], which produced many valuable insights about the Continuous Double Auction (CDA) mechanism, and strategies for bidding in CDAs. Even though the CDA had been well studied before, the focused effort and attention catalyzed by the competition led to substantial cross-fertilization of ideas and perspectives. The book generated from that event [Friedman and Rust, 1993] remains a seminal reference for CDAs.

A multi-year event like TAC offers the further prospect of learning from shared experience. A goal of repeating the competition was to observe the progress of trading agents, in effect accelerating the evolution of an adapted population of traders. Now that we have three years of experience in the TAC series, it is appropriate to examine whether this objective has been fulfilled. Moreover, given all the effort devoted to TAC participation, it is incumbent on us to exploit the TAC data for what they are worth, and see what conclusions we might draw about the efficacy of trading agents and particular ideas about trading strategy.

By one measure, TAC has certainly succeeded in spurring research. Over a dozen publications reporting on the competitions, specific agents, techniques employed, and analyses have appeared to date in archival journals, refereed conferences, and magazines.[1] The TAC "literature" thus represents an uncommonly rich corpus of documentation on trading strategy and behavior for a particular complex environment. Many of the accounts include specific analyses or experiments involving agents from multiple developers, or variants on a particular agent inspired by techniques reportedly employed by others [Greenwald, 2003b; Stone *et al.*, 2002; Wellman *et al.*, 2002]. Such efforts augment the anecdotal evidence from entrants that each successive year the lessons and approaches presented previously are incorporated in new and improved agent designs.

In this paper, we present some data bearing on the assessment of performance and progress of trading agents, as reflected in the TAC series to date. We employ data from the actual TAC tournaments, as well as some post-competition experimentation. Our analysis is based almost entirely on outcomes (profits and allocations), with very little direct accounting for specific agent techniques. Though we offer some conclusions, our investigation also raises further questions. A definitive assessment of agent competence must await further studies, perhaps based on succeeding years of TAC experience.

Our presentation assumes the reader is familiar with the general structure of the TAC travel shopping game. A complete description is available on the TAC web site (`http://tac.eecs.umich.edu/`), as well as many of the cited papers.

2 TAC-02 Tournament Results

Average scores for the sixteen agents that played in the final and semifinal rounds are posted in Table 1. See `http://www.sics.se/tac` for a list of participant affiliations and team leaders, as well as results from preliminary rounds. Complete game logs are available, as they are for the previous TAC events. Brief agent descriptions have been collected by Greenwald [2003a].

Although we agree with those who have cautioned against focusing excessively on ranked results in the context of research competitions [Stone, 2002], tournament results provide an important source of information about agent quality. Agents are presumed to act to maximize expected score, and so all else equal, an increase in score reflects an improved agent. If several agents improve, however, this may or may not lead to higher total scores for those agents. Whereas some kinds of improvement unambiguously increase total agent

[1] `http://auction2.eecs.umich.edu/researchreport.html`

Agent	Affiliation	Scores		
		Semifinals	Finals	CP Adj
ATTac	AT&T Research (et al.)	H1: 3137	—	—
cuhk	Chinese U Hong Kong	H2: 3266	3069	-24
kavayaH	Oracle India	H1: 3200	3099	-60
livingagents[2]	Living Systems AG	H1: 3310	3181	-20
PackaTAC	N Carolina State U	H2: 3250	—	—
PainInNEC	NEC Research (et al.)	H1: 2193	—	—
RoxyBot	Brown U	H2: 3160	—	—
sics	Swedish Inst Comp Sci	H2: 3146	—	—
SouthamptonTAC	U Southampton	H1: 3397	3385	-48
Thalis	U Essex	H2: 3199	3246	-36
tniTac	Poli Bucharest	H1: 3108	—	—
TOMAhack	U Toronto	H2: 2843	—	—
tvad	Technion	H1: 2724	—	—
UMBCTAC	U Maryland Baltimore Cty	H1: 3208	3236	+55
Walverine	U Michigan	H2: 3287	3210	+67
whitebear	Cornell U	H2: 3324	3413	+66

Table 1: TAC-02 seeded agents, and their average scores during the semifinals (14 games) and finals (32 games). The third column represents our calculated adjustment to final-round scores due to client preference assignments.

surplus (e.g., fewer wasted flights, better allocation of entertainment), others may reduce the value retained by agents (e.g., smarter agents may be more effective at competing away the consumer surplus) or even the total system surplus (e.g., deadweight loss due to strategic behavior).

One way to measure progress over time is to track benchmark levels of performance by keeping some agents constant. For example, in the CADE ATP (automated theorem proving) series of competitions [Sutcliffe, 2001], the best systems from a given year typically enter unchanged in the next year's event (along with improved versions, of course). This provides a direct measure, in comparable terms, of the relative performance across years. In a game setting, where other agents are part of the environment, it is not strictly fair to judge an agent with respect to a different field. Nevertheless, it can be quite instructive to observe the implications of such transplants.

The 2001 tournament [Wellman *et al.*, 2003] included two calibrating agents in the seeding round. ATTac-2000 [Stone *et al.*, 2001] represented the highest-scoring agent from the TAC-00 finals. To account for the rule changes between TAC-00 and TAC-01, ATTac-2000 was modified with a one-line edit causing it to place all of its bids before the first hotel closure as opposed to during the last minute of the game. We also included dummy_buyer, the agent provided by the Michigan TAC team in 2001 to play in test games that do not have a full slate of agents. Whereas most of the other agents' behaviors were modified between (and during) the qualifying and seeding round, the dummy was left unchanged. Not surprisingly, we observed substantial deterioration in the dummy's standing as the preliminary rounds progressed.

The 2002 tournament did not explicitly insert calibrating agents, but the fact that twelve of the participating teams from 2001 also entered agents in TAC-02 provides some natural calibration. In particular, the two top-scoring agents in TAC-01, livingagents [Fritschi and Dorer, 2002] and ATTac-2001, participated with essentially unchanged agents in TAC-02. As shown in the table, livingagents did quite well, assuming we ignore the bug that caused it to skip two games. ATTac was top scorer in the TAC-02 seeding rounds, but then was eliminated in the semifinals. One possible explanation is simply that the agent experienced technical difficulties due to a change of computational environments between the seeding and semifinal rounds.[3] Another more substantive possibility is that prices during the preliminary rounds in 2002 (which ATTac uses as training data) were not sufficiently representative of the final rounds. We suspect that the decrease in relative performance also reflects a general increase in competence of the other agents in the field. Interestingly, it may well be that because livingagents in some respects benefits by playing along with effective and adaptive agents [Wellman *et al.*, 2003], it may be more robust with respect to improvements in the rest of the field.

The two top-scoring agents in TAC-02, whitebear and SouthamptonTAC [He and Jennings, 2002], also contended in TAC-01. These agents reportedly evolved from their 2001 designs, improved through adopting refined classifications of game environments [He and Jennings, 2003], and through extensive experimentation and parameter tuning [Vetsikas and Selman, 2003].

Of the eight other repeat entries:

- Three represent complete reimplementations of the corresponding TAC-01 entries, by essentially different agent designers (Thalis, sics, UMBCTAC).
- Two represent significant redesigns by the same essen-

[2] The score of livingagents was adversely affected by missing two games. Discounting these would have led to an average score of 3393.

[3] Peter Stone, personal communication.

tial designers (RoxyBot, cuhk).

- Three represent incremental or unknown changes by the same or related designers (PainInNEC, BigRed, harami).

3 TAC Market Efficiency

Another gauge of agent effectiveness is how well they allocate travel goods, *in the aggregate*, through their market interactions. This is an indirect measure, at best, since the objective of each agent is to maximize its own surplus, not that of the overall system—comprising all agents plus the TAC seller. Nevertheless, such a *social welfare* analysis can provide a benchmark, and shed light on the allocation of resources through an economy of interacting software agents.

3.1 Market Efficiency in the TAC-02 Tournament

We can measure aggregate effectiveness by comparing actual TAC market allocations with ideal global allocations. Consider the total group of 64 clients, and the set of available resources: 16 hotel rooms of each type per day, plus 8 entertainment tickets of each type per day. The global optimizer calculates the allocation of resources maximizing total client utility, net of expenditures on flights assuming they are available at their initial prices. We take initial prices to be the relevant inherent cost (exogenously determined, independent of TAC agent demand) of flights, treating the expected stochastic increase of flights during the game as a cost of decision delay that would be avoided by the idealized optimizer. Note that the global optimization completely neglects hotel and entertainment prices, as these are endogenous to the TAC market. Monetary transfers affect the distribution of surplus across TAC buyers and sellers, but not the total amount. We formulate the optimization problem as an integer linear program, and solve it using CPLEX.

The average achievable net utility, per client, in the various rounds of the TAC tournament as determined by global optimization is reported under the heading “Global” in Table 2. Average net utility achieved in the actual TAC games (also neglecting hotel and entertainment expenditures, but counting actual payments for flights) is reported under “TAC Market”.

Round	Games	Global	TAC Market	TAC (%)
Qualify	390	618	415	67.0
Seeding	1045	618	470	75.7
Semi-Final	28	608	534	87.7
Final	32	609	542	89.1

Table 2: The efficiency of the TAC market compared to the global optimum.

As seen in the table, we found that the TAC market achieved 89% of the optimal value, on average, over the 32 games of the TAC-02 finals. There was a steady improvement from the qualifying round (67% optimal), seeding round (76%), and semifinals (88%). All of these differences are significant ($p < 0.01$), except the small increment from semifinals to finals.[4]

It is difficult to assess this effectiveness in absolute terms, so we provide a couple of benchmarks for comparison.

1. Uniform hotel and entertainment. We distribute the hotel rooms and entertainment evenly across the eight agents, then optimize each agent’s allocation to clients. This approach yields 95.2% of the globally optimal value on average. (Allocation values were significantly better than the market in every round.)
2. Uniform hotel, endowed entertainment. The relative average value drops to 85.4% if we distribute only the hotels, leaving agents with their original endowment of entertainment. (This value represents a significant improvement to the market in qualifying and seeding rounds, with the market significantly better in the finals.)

It is perhaps surprising that simply dividing the goods uniformly achieves such a high fraction of the available surplus—better than the market if entertainment is included in the distribution. One reason that the uniform distribution is relatively so effective is that the agents are *ex ante* symmetric, with i.i.d. clients. Potential gains from trade are thus not so great for hotels. Second, a direct allocation avoids the significant obstacles posed to agents pursuing their allotments individually through the market. Agents face substantial risk (price uncertainty, exposure due to complementarities, unknown hotel closing patterns), and this necessarily entails some loss in expected allocation quality. For example, the set of available hotels is sufficient to obtain trips for all clients (albeit shortened from desired lengths), and given a definite allocation the agent can optimize for its clients accordingly. With uncertainty, the agents may plan for longer trips than are jointly feasible, and thus wind up wasting flights, hoarding hotel rooms (to hedge), or resorting to suboptimal fallback trip options. In future work, we will investigate in greater depth the various sources of misallocation in TAC play.

3.2 Comparisons

Given that agent programmers are actively debugging and developing their agents during the preliminary rounds, it seems fair to assume that agent competence improves in succeeding rounds of the tournament. The selection of best performers for the semi-finals and then the finals naturally amplifies this effect. Thus, the progressive improvement in market efficiency observed in Table 2 coincides with individual agent progress. We performed the same global optimization analysis for the TAC-01 finals (24 games), and found a market efficiency of 85.7%. Though better than the TAC-02 qualifying and seeding rounds, the TAC-01 finalists did not allocate resources as well as TAC-02 finalists ($p = 0.024$), or even the TAC-02 semi-finalists ($p = 0.097$). This indirectly (to the extent that overall market efficiency aligns with individual agent success) confirms our conjecture that the 2002 agents were on the whole more competent than their predecessors.

It is also potentially interesting to compare market effectiveness across different configurations of agents. Our initial

[4] Henceforth, all assertions of statistical significance are with respect to the 0.01 level, unless otherwise specified.

explorations employ versions of Michigan's TAC-02 entry, Walverine [Cheng *et al.*, 2003]. In 54 games with Walverine playing all eight agent slots, the market achieved 89.8% efficiency, a result not statistically distinguishable from that of the actual pool of TAC-02 finalists.

Of course, Walverine, like the others, aims to promote its own profit, not overall efficiency per se. Indeed, we can identify particular strategy components that would be *expected* to detract from social welfare. Specifically, Walverine *shades* its hotel bids, generally offering a price strictly lower than its actual marginal value for a given hotel room. It determines bid prices as part of a decision-theoretic optimization of expected surplus, which takes into account the probability of not obtaining a good even though its price is below the agent's valuation [Cheng *et al.*, 2003]. Though it benefits the individual agent by design, such shading runs counter to the goal of market efficiency, as the market does not generally have available faithful signals of the relative value of goods to the various agents.[5]

To evaluate this effect, we ran another trial of 60 games played by Walverine variants with their optimal bidding procedure (i.e., shading) turned off. Agents in this version bid their true marginal values for every hotel room. On average, the market with non-shading Walverines achieved 91.3% efficiency, significantly better ($p = 0.051$) than the actual TAC-02 finals.

In further work, we intend to evaluate additional agent configurations, including further Walverine variants as well as agents developed by other groups.[6] One interesting question (at least as another benchmark) is what level of overall efficiency can be attained by agents actually designed with this objective in mind. It should be possible to achieve at least the level of our uniform-hotel-and-entertainment benchmark (95.2%), since it is straightforward to construct bidding policies that implement a uniform distribution of all goods.

3.3 Entertainment Trading Efficiency

One component of market effectiveness amenable to separate analysis is entertainment trading. Entertainment goods are initially distributed as endowments to the agents, who exchange among themselves through CDAs to reach a final allocation. Although the value of entertainment to agents depends on their choice of trip dates, it is possible to characterize with reasonable accuracy the gains from trade specifically attributable to the entertainment component of the TAC market.

To measure entertainment trading efficiency, we simply compare the aggregate "fun bonus" component of trip utility in the globally optimal allocation, with that attained in the actual TAC market. Efficiency percentages for the various game sets are presented in Table 3, which also repeats the overall efficiency numbers for convenient reference.

One surprising observation is that the entertainment trading performance for the non-shading Walverine is significantly better than that of the unmodified Walverine, despite the fact that their entertainment trading policies are identical (shading applies only to hotel bids). We surmise that the difference must be due to the superior allocation of hotels in the non-shading case. Since hotel valuations reflect entertainment opportunities, better hotel allocations entail better use of entertainment goods.

Another interesting result is that the entertainment performance in the TAC-02 finals was virtually the same as in the TAC-01 finals, despite the significant improvement in market performance overall. This suggests that the strategic progress was focused on hotel and flight strategies, which certainly agrees with the entrants' reports of their concentrations of effort.

The data on entertainment performance can also help to calibrate estimates of potential gains from agent improvements. We reported above on two benchmarks based on uniform hotel distributions, one with no entertainment trading and one with uniform entertainment allocation. These define a potential gain from trade of approximately 60.5 per client (484 per agent) given the fixed uniform allocation of hotels. In our process of training Walverine's entertainment strategy [Cheng *et al.*, 2003], we observed a difference of roughly 478 on average between a policy of not trading entertainment (average fun bonus 1019), and that of a representative hand-coded strategy (that of livingagents, average fun bonus 1497). In contrast, the average fun bonus in the global optimal allocation runs around 1677. The fact that the observed gain from trade in training[7] approaches the gain from uniform entertainment distribution in the uniform-hotel case suggests that the remaining benefit from entertainment trading is equal to the difference between uniform and optimal. We evaluated this by calculating a third benchmark, based on global optimization of entertainment subject to a uniform allocation of hotels to agents. The result is 92.7 per agent greater than the value with uniform entertainment allocation.

4 Comment

The foregoing analysis provides some evidence for competence and progress in TAC traders. Since our measures are all indirect (e.g., measuring market efficiency rather than absolute agent performance), however, definitive conclusions are not justified. More compelling demonstrations of progress and competence might be based on further calibration studies, systematic search in strategy spaces, and attribution of allocation suboptimality among its many possible causes (e.g., agent suboptimality, and inherent risk—including the cost of its rational management). Further benchmarks, capturing less ideal conditions, may prove useful in this regard.

[5] If all agents shaded proportionally, then the relative offer prices would still provide the relevant information to the market. In general, however, the price reductions do not cancel out in this way, as the bidder's optimization includes many agent-specific contextual factors.

[6] Performing our analysis requires only data describing client preferences and final allocations. We welcome any game logs other researchers would be willing to submit for our efficiency analysis.

[7] The fun bonus realized in the TAC-02 finals turned out to be somewhat lower than expected based on training observations, for both Walverine and livingagents, and apparently the rest of the field as well except for whitebear [Cheng *et al.*, 2003].

Round	Games	TAC (%)	Entertainment (%)
02 Qualify	390	67.0	71.1
02 Seeding	1045	75.7	79.0
02 Semi-Final	28	87.7	83.0
02 Final	32	89.1	85.3
01 Final	24	85.7	85.5
all Walverine	54	89.8	83.8
non-shading Walverine	60	91.3	86.1

Table 3: The efficiency of the TAC market compared to the global optimum—overall and specifically with respect to entertainment.

Another natural question not at all addressed by this work is how well TAC agents fare compared to what human traders could do? We are aware of no evidence that humans would be particularly adept at a TAC-like trading task. One of the few studies comparing human and computer traders (in an abstract CDA scenario) did not reflect very favorably on the humans [Das *et al.*, 2001].

Acknowledgments

We benefited much from discussions with the rest of the Michigan TAC Team. This research was supported in part by NSF grants IIS-9988715 and IIS-0205435.

References

[Cheng *et al.*, 2003] Shih-Fen Cheng, Evan Leung, Kevin M. Lochner, Kevin O'Malley, Daniel M. Reeves, and Michael P. Wellman. Walverine: A Walrasian trading agent. In *Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Melbourne, 2003.

[Das *et al.*, 2001] Rajarshi Das, James E. Hanson, Jeffrey O. Kephart, and Gerald Tesauro. Agent-human interactions in the continuous double auction. In *Seventeenth International Joint Conference on Artificial Intelligence*, pages 1169–1176, Seattle, WA, 2001.

[Friedman and Rust, 1993] Daniel Friedman and John Rust, editors. *The Double Auction Market*. Addison-Wesley, 1993.

[Fritschi and Dorer, 2002] Clemens Fritschi and Klaus Dorer. Agent-oriented software engineering for successful TAC participation. In *First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Bologna, 2002.

[Greenwald, 2003a] Amy Greenwald. The 2002 trading agent competition: An overview of agent strategies. *AI Magazine*, 24(1):83–91, 2003.

[Greenwald, 2003b] Amy Greenwald. Bidding under uncertainty in simultaneous auctions. Technical report, Brown University, 2003.

[He and Jennings, 2002] Minghua He and Nicholas R. Jennings. SouthamptonTAC: Designing a successful trading agent. In *Fifteenth European Conference on Artificial Intelligence*, pages 8–12, Lyon, 2002.

[He and Jennings, 2003] Minghua He and Nicholas R. Jennings. SouthamptonTAC: An adaptive autonomous trading agent. *ACM Transactions on Internet Technology*, 2003.

[Rust *et al.*, 1993] John Rust, John Miller, and Richard Palmer. Behavior of trading automata in a computerized double auction market. In Friedman and Rust [1993], pages 155–198.

[Stone *et al.*, 2001] Peter Stone, Michael L. Littman, Satinder Singh, and Michael Kearns. ATTac-2000: An adaptive autonomous bidding agent. *Journal of Artificial Intelligence Research*, 15:189–206, 2001.

[Stone *et al.*, 2002] Peter Stone, Robert E. Schapire, János A. Csirik, Michael L. Littman, and David McAllester. ATTac-2001: A learning, autonomous bidding agent. In *Agent-Mediated Electronic Commerce IV*, volume 2153 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.

[Stone, 2002] Peter Stone. Multiagent competitions and research: Lessons from RoboCup and TAC. In *Sixth RoboCup International Symposium*, Fukuoka, Japan, 2002.

[Sutcliffe, 2001] Geoff Sutcliffe. The CADE-17 ATP system competition. *Journal of Automated Reasoning*, 27:227–250, 2001.

[Vetsikas and Selman, 2003] Ioannis A. Vetsikas and Bart Selman. A principled study of the design tradeoffs for autonomous trading agents. In *Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Melbourne, 2003.

[Wellman and Wurman, 1999] Michael P. Wellman and Peter R. Wurman. A trading agent competition for the research community. In *IJCAI-99 Workshop on Agent-Mediated Electronic Trading*, Stockholm, August 1999.

[Wellman *et al.*, 2002] Michael P. Wellman, Daniel M. Reeves, Kevin M. Lochner, and Yevgeniy Vorobeychik. Price prediction in a trading agent competition. Technical report, University of Michigan, 2002.

[Wellman *et al.*, 2003] Michael P. Wellman, Amy Greenwald, Peter Stone, and Peter R. Wurman. The 2001 trading agent competition. *Electronic Markets*, 13:4–12, 2003.