Optimal Payoff Functions for Members of Collectives

David H. Wolpert
NASA Ames Research Center
Moffett Field, CA 94035
dhw@ptolemy.arc.nasa.gov

Kagan Tumer
NASA Ames Research Center
Moffett Field, CA 94035
kagan@ptolemy.arc.nasa.gov

September 20, 2001

Abstract

We consider the problem of designing (perhaps massively distributed) collectives of computational processes to maximize a provided "world" utility function. We consider this problem when the behavior of each process in the collective can be cast as striving to maximize its own payoff utility function. For such cases the central design issue is how to initialize/update those payoff utility functions of the individual processes so as to induce behavior of the entire collective having good values of the world utility. Traditional "team game" approaches to this problem simply assign to each process the world utility as its payoff utility function. In previous work we used the "Collective Intelligence" (COIN) framework to derive a better choice of payoff utility functions, one that results in world utility performance up to orders of magnitude superior to that ensuing from use of the team game utility. In this paper we extend these results using a novel mathematical framework. We review the derivation under that new framework of the general class of payoff utility functions that both i) are easy for the individual processes to try to maximize, and ii) have the property that if good values of them are achieved, then we are assured of a high value of world utility. These are the "Aristocrat Utility" and a new variant of the "Wonderful Life Utility" that was introduced in the previous COIN work. We demonstrate experimentally that using these new utility functions can result in significantly improved performance over that of previously investigated COIN payoff utilities, over and above those previous utilities' superiority to the conventional team game utility. These results also illustrate the substantial superiority of these payoff functions to the perhaps the most natural version of the economics technique of "endogenizing externalities".

1 Introduction

In this paper we are concerned with large distributed collectives of interacting goal-driven computational processes, where there is a provided 'world utility' function that rates the possible behaviors of that collective [31, 30]. We are particularly concerned with such collectives where the individual computational processes use machine learning techniques (e.g., Reinforcement Learning (RL) [14, 22, 21, 25]) to try to achieve their individual goals. We represent those goals of the individual processes as maximizing an associated 'payoff' utility function, one that in general can differ from the world utility.

In such a system, we are confronted with the following inverse problem: *How should one initialize/update the payoff utility functions of the individual processes so that the ensuing behavior of the entire collective achieves large values of the provided world utility?* In particular, since in truly large systems detailed modeling of the system is usually impossible, how can we avoid such modeling? Can we instead leverage the simple assumption that our learnering algorithms are individually fairly good at what they do to achieve a large world utility value?

This problem is related to work in many other fields, including multi-agent systems (MAS's), computational economics, mechanism design, reinforcement learning, statistical mechanics, computational ecologies, (partially observable) Markov decision processes and game theory. However none of these fields is both applicable in large problems, and directly addresses the *general* inverse problem, rather than a special instance of it. (See [30] for a detailed discussion of the relationship between these fields, involving hundreds of references.) For example, the field of mechanism design is not generally applicable, being largely tailored to collectives of human beings, and in particular to the idiosyncracy of such collectives that their members have hidden variables whose values they "do not want to reveal". There is other previous work that does consider the general inverse problem, and even has each individual computational process (or "agent") use reinforcement learning [2, 7, 10, 15, 18]. However, in that work in general each process has the world utility function as its payoff utility function (i.e., implements a "team game" or an "exact potential game" [8]). Unfortunately, as expounded below and in previous work, this approach scales extremely poorly to large problems. (Intuitively, the difficulty is that each agent can have a hard time discerning the echo of its behavior on the world utility when the system is large; each agent has a horrible "signal-to-noise" problem.)

Intuitively, we are concerned with payoff utility functions that are "aligned" with the world utility, in that modifications a player might make that would improve its payoff utility also must improve world utility. Fortunately the equivalence class of such payoff utilities exends well beyond team-game utilities. As a particular example, in previous work we used the COllective INtelligence (COIN) framework to derive the 'Wonderful Life Utility' (WLU) payoff function [30] as an alternative to a team-game payoff utility. The WLU is aligned with world utility, as desired. In addition though, WLU overcomes much of the signal-to-noise problem of team game utilities [24, 31, 30, 33].

As an example, in some of our previous work we used the WLU for distributed control of network packet routing [31]. Conventional approaches to packet routing have each router run a shortest path algorithm (SPA), i.e., each router routes its packets in the way that it expects will get those packets to their destinations most quickly. Unlike with a WLU-based collective, with SPA-based routing the routers have no concern for the possible deleterious side-effects of their routing decisions on the global goal (e.g., they have no concern for whether they induce bottlenecks). We ran simulations that demonstrated that a WLU-based collective has substantially better throughputs than does the best possible SPA-based system [31], even though that SPA-based system has information denied the COIN system. In related work we have shown that use of the WLU automatically avoids the infamous Braess' paradox, in which adding new links can actually decrease throughput — a situation that readily ensnares SPA's.

As another example, in [32] we considered the pared-down problem domain of a congestion game, in particular a more challenging variant of Arthur's El Farol bar attendance problem [1], sometimes also known as the "minority game" [6]. In this problem the individual processes making up the collective are explicitly viewed as 'players' involved in a non-cooperative game. Each player has to determine which night in the week to attend a bar. The problem is set up so that if either too few people attend (boring evening) or too many people attend (crowded evening), the total enjoyment of the attending players drops. Our goal is to design the payoff functions of the players so that the total enjoyment across all nights is maximized. In this previous work we showed that use of the WLU can result in performance *orders of magnitude* superior to that of team game utilities.

In this article we extend this previous work, by investigating the impact of the choice of the single free parameter in the WLU (the 'clamping parameter'), which we simply set to 0 in our previous work. In particular, we employ some of the new mathematics of COINs to derive the 'Aristocrat Utility' (AU) as an optimal utility payoff function. We derive the optimal value of the clamping parameter as the value that gives a "mean-field" approximation to the AU. We then present experimental tests to validate that choice of clamping parameter. In the next section we review the relevant concepts of the mathematics of COINs. Then we sketch how to use those concepts to derive the optimal clamping parameter. To facilitate comparison with previous work, we chose to conduct our experimental investigations of the performance with this optimal clamping parameter in variations of the Bar Problem. We present those variations in Section 3. Finally we present the results of the experiments in Section 4. Those results corroborate the predicted improvement in performance when using our theoretically derived clamping parameter. This extends the superiority of the COIN-based approach above conventional team-game approaches even further than had been done previously. The results also illustrate the substantial superiority of COIN-based techniques to a natural version of the economics technique of "endogenizing externalities".

2 The Mathematics of Collective Intelligence

Since in this paper we are restricting attention to variants of the bar problem, we view the individual computational processes as players involved in an iterated single-stage game. The full mathematics of the COIN framework extends significantly beyond what is needed to address such games [1] The restricted version we will call upon here starts with an arbitrary vector space Z whose elements ζ give the joint move of all players in the collective in some stage. We wish to search for the ζ that maximizes the provided **world utility** $G(\zeta)$. In addition to G we are concerned with **payoff utility functions** $\{g_\eta\}$, one such function for each variable/player η. We use the notation $\hat{\eta}$ to refer to all players other than η.

We will need to have a way to "standardize" utility functions so that the numeric value they assign to a ζ only reflects their ranking of ζ relative to certain other elements of Z. We call such a standardization of some arbitrary utility U for player η the "**intelligence** for η at ζ with respect to U". Here we will use intelligences that are equivalent to percentiles:

[1] That framework encompasses, for example, arbitrary dynamic redefinitions of the "players" (i.e., dynamic reassignments of how the various subsets of the variables comprising the collective are assigned to players), as well as modification of the players' information sets (i.e., modification of inter-player communication). See [28].

$$\epsilon_U(\zeta : \eta) \equiv \int d\mu_{\zeta_{,\eta}}(\zeta')\Theta[U(\zeta) - U(\zeta')] \ , \tag{1}$$

where the Heaviside function Θ is defined to equal 1 when its argument is greater than or equal to 0, and to equal 0 otherwise, and where the subscript on the (normalized) measure $d\mu$ indicates it is restricted to ζ' sharing the same non-η components as ζ.[2] Intelligence value are always between 0 and 1.

Our uncertainty concerning the behavior of the system is reflected in a probability distribution over Z. Our ability to control the system consists of setting the value of some characteristic of the collective, e.g., setting the payoff functions of the players. Indicating that value by s, our analysis revolves around the following central equation for $P(G \mid s)$, which follows from Bayes' theorem:

$$P(G \mid s) = \int d\vec{\epsilon}_G P(G \mid \vec{\epsilon}_G, s) \int d\vec{\epsilon}_g P(\vec{\epsilon}_G \mid \vec{\epsilon}_g, s) P(\vec{\epsilon}_g \mid s) \ , \tag{2}$$

where $\vec{\epsilon}_g \equiv (\epsilon_{g_{\eta_1}}(\zeta : \eta_1), \epsilon_{g_{\eta_2}}(\zeta : \eta_2), \cdots)$ is the vector of the intelligences of the players with respect to their associated payoff functions, and $\vec{\epsilon}_G \equiv (\epsilon_G(\zeta : \eta_1), \epsilon_G(\zeta : \eta_2), \cdots)$ is the vector of the intelligences of the players with respect to G.

Note that $\epsilon_{g_\eta}(\zeta : \eta) = 1$ means that player η is fully rational at ζ, in that its move maximizes its payoff, given the moves of the players. In other words, a point ζ where $\epsilon_{g_\eta}(\zeta : \eta) = 1$ for all players η is one that meets the definition of a game-theory Nash equilibrium.[3] On the other hand, a ζ at which all components of $\vec{\epsilon}_G = 1$ is a local maximum of G (or more precisely, a critical point of the $G(\zeta)$ surface).

If we can choose s so that the third conditional probability in the integrand is peaked around vectors $\vec{\epsilon}_g$ all of whose components are close to 1, then we have likely induced large (payoff function) intelligences. If we can also have the second term be peaked about $\vec{\epsilon}_G$ equal to $\vec{\epsilon}_g$, then $\vec{\epsilon}_G$ will also be large. Finally, if the first term in the integrand is peaked about high G when $\vec{\epsilon}_G$ is large, then our choice of s will likely result in high G, as desired.

Intuitively, the requirement that payoff functions have high "signal-to-noise" (an issue not considered in conventional work in mechanism design) arises in the third term. It is in the second term that the requirement that the payoff functions be "aligned with G" arises. In this work we concentrate on these two terms, and show how to simultaneously set them to have the desired form. [4]

Details of the stochastic environment in which the collective operates, together with details of the learning algorithms of the players, are reflected in the distribution $P(\zeta)$ which

[2]The measure must reflect the type of system at hand, e.g., whether Z is countable or not, and if not, what coordinate system is being used. Other than that, any convenient choice of measure may be used and the theorems will still hold.

[3]See [9]. Note that consideration of points ζ at which not all intelligences equal 1 provides the basis for a model-independent formalization of bounded rationality game theory, a formalization that contains variants of many of the theorems of conventional full-rationality game theory. See [27].

[4]Non-game-theory-based function maximization techniques like simulated annealing instead address how to have term 1 have the desired form. They do this by trying to ensure that the local maxima that the underlying system ultimately settles near have high G, by "trading off exploration and exploitation". One can combine such term-1-based techniques with the techniques presented here, The resultant hybrid algorithm, addressing all three terms, outperforms simulated annealing by over two orders of magnitude[29].

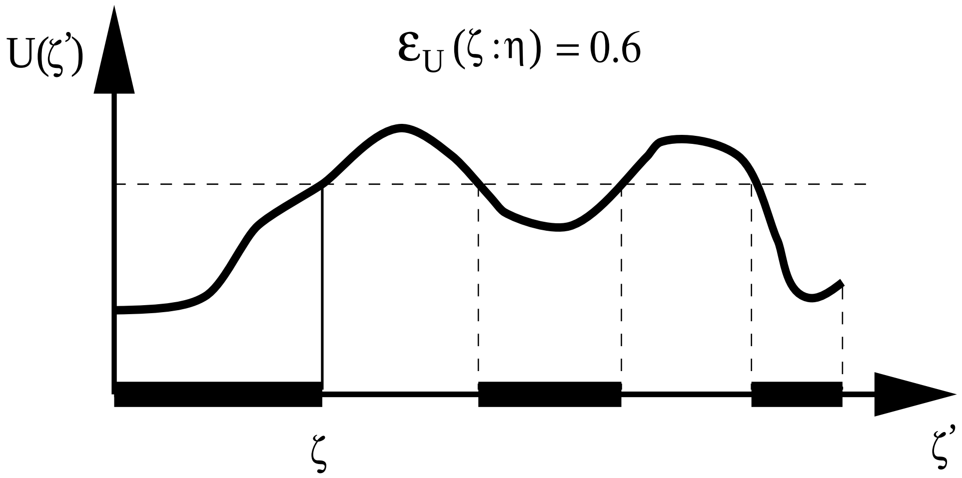

Figure 1: Intelligence of agent η at state ζ for utility U: ζ is the actual joint move at hand. The x-axis shows agent η's alternative possible moves (all states ζ' having ζ's values for the moves of all players other than η.). The thick sections of the x-axis show the alternative moves that η could have made that would have given η a worse value of the utility U. The fraction of the full set of η's possible moves that lies in those thick sections (which is 0.6 in this example) is the intelligence of agent η at ζ for utility U, denoted by $\epsilon_U(\zeta : \eta)$.

underlies the distributions appearing in Equation 2. Note though that *independent of these considerations*, our desired form for the second term in Equation 2 is assured if we have chosen payoff utilities such that $\vec{\epsilon}_g$ equals $\vec{\epsilon}_G$ exactly *for all* ζ. We call such a system *factored*. In game-theory language, the Nash equilibria of a factored collective are local maxima of G. In addition to this desirable equilibrium behavior, factored collectives also automatically provide appropriate off-equilibrium incentives to the players (an issue rarely considered in the game theory / mechanism design literature).

As a trivial example, any "team game" in which all the payoff functions equal G is factored [8, 16]. However team games often have very poor forms for term 3 in Equation 2, forms which get progressively worse as the size of the collective grows. This is because for such payoff functions each player η will usually confront a very poor "signal-to-noise" ratio in trying to discern how its actions affect its payoff $g_\eta = G$, since so many other player's actions also affect G and therefore dilute η's effect on its own payoff function.

We now focus on algorithms based on payoff functions $\{g_\eta\}$ that optimize the signal/noise ratio reflected in the third term, subject to the requirement that the system be factored. To understand how these algorithms work, given a measure $d\mu(\zeta_\eta)$, define the **opacity** at ζ of utility U as:

$$\Omega_U(\zeta : \eta, s) \equiv \int d\zeta' J(\zeta' \mid \zeta) \frac{|U(\zeta) - U(\zeta'_{\hat{\eta}}, \zeta_\eta)|}{|U(\zeta) - U(\zeta_{\hat{\eta}}, \zeta'_\eta)|} \; , \tag{3}$$

where J is defined in terms of the underlying probability distributions,[5] and $(\zeta'_{\hat{\eta}}, \zeta_\eta)$ is defined as the worldline whose $\hat{\eta}$ components are the same as those of ζ' while its η components are

[5]Writing it out in full, $J(\zeta' \mid \zeta) \equiv J(\zeta_\eta, \zeta' \mid \zeta_{\hat{\eta}}, s)/P(\zeta_\eta \mid \zeta_{\hat{\eta}}, s)$, with:

$$J(\zeta_\eta, \zeta' \mid \zeta_{\hat{\eta}}, s) \equiv \frac{P(\zeta_\eta \mid \zeta_{\hat{\eta}}, s) P(\zeta'_\eta \mid \zeta_\eta, s)\mu(\zeta'_\eta)}{2} + \frac{P(\zeta'_\eta \mid \zeta'_{\hat{\eta}}, s) P(\zeta_\eta \mid \zeta'_\eta, s)\mu(\zeta_\eta)}{2} \; .$$

the same as those of ζ.

The denominator absolute value in the integrand in Equation 3 reflects how sensitive $U(\zeta)$ is to changing ζ_η. In contrast, the numerator absolute value reflects how sensitive $U(\zeta)$ is to changing $\zeta_{\hat{\eta}}$. So the smaller the opacity of a payoff function g_η, the more $g_\eta(\zeta)$ depends only on the move of player η, i.e., the better the associated signal-to-noise ratio for η. Intuitively then, lower opacity should mean it is easier for η to achieve a large value of its intelligence.

To formally establish this, we use the same measure $d\mu$ to define opacity as the one that defined intelligence. Under this choice expected opacity bounds how close to 1 expected intelligence can be [28]:

$$
\begin{aligned}
&E(\epsilon_U(\zeta : \eta) \mid s) \leq 1 - K, \text{where} \\
&K \leq E(\Omega_U(\zeta : \eta, s) \mid s).
\end{aligned} \tag{4}
$$

So low expected opacity of utility g_η ensure that a necessary condition is met for the third term in Equation 2 to have the desired form for player η. While low opacity is not, formally speaking, also sufficient for $E(\epsilon_U(\zeta : \eta) \mid s)$ to be close to 1, in practice the bounds in Equation 4 are usually tight.

It is possible to solve for the set of all payoff utilities that are factored with respect to a particular world utility. Unfortunately, in general it is not possible for a collective both to be factored and to have zero opacity for all of its players. However consider **difference** utilities, which are of the form

$$
U(\zeta) = G(\zeta) - \Gamma(f(\zeta)) \tag{5}
$$

where $\Gamma(f)$ is independent ζ_η. Any difference utility is factored [28]. In addition, under usually benign approximations, $E(\Omega_u \mid s)$ is minimized over the set of difference utilities by choosing

$$
\Gamma(f(\zeta)) = E(G \mid \zeta_{\hat{\eta}}, s) \ , \tag{6}
$$

up to an overall additive constant. We call the resultant difference utility the **Aristocrat** utility (AU), loosely reflecting the fact that it measures the difference between a player's actual action and the average action.

If possible, we would like each player η to use the associated AU as its payoff function to ensure good form for both terms 2 and 3 in Equation 2. This is not always feasible however. The problem is that to evaluate the expectation value defining its AU each player needs to evaluate the current probabilities of each of its potential moves. However if the player then changes its payoff function to be the associated AU it will in general substantially change its ensuing behavior. (The player now wants to choose moves that maximize a different function from the one it was maximizing before.) In other words, it will change the probabilities of its moves, which means that its new payoff function is in fact not the AU for its actual (new) probabilities.

There are ways around this self-consistency problem, but in practice it is often easier to bypass the entire issue, by giving each η a payoff function that does not depend on the

$$
\begin{array}{c}
\\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4
\end{array}
\begin{array}{c}
\zeta \\
\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}
\end{array}
\quad
\begin{array}{c} \Longrightarrow \\ \text{Clamp } \eta_2 \\ \text{to "null"} \end{array}
\quad
\begin{array}{c}
(\zeta_{\hat{\eta}_2}, \vec{0}) \\
\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}
\end{array}
\quad
\begin{array}{c} \Longrightarrow \\ \text{Clamp } \eta_2 \\ \text{to "average"} \end{array}
\quad
\begin{array}{c}
(\zeta_{\hat{\eta}_2}, \vec{a}) \\
\begin{bmatrix} 1 & 0 & 0 \\ .33 & .33 & .33 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}
\end{array}
$$

Figure 2: This example shows the impact of the clamping operation on the joint state of a four-player system where each player has three possible moves, each such move represented by a three-dimensional unary vector. The first matrix represents the joint state of the system ζ where player 1 has selected action 1, player 2 has selected action 3, player 3 has selected action 1 and player 4 has selected move 2. The second matrix displays the effect of clamping player 2's action to the "null" vector (i.e., replacing ζ_{η_2} with $\vec{0}$). The third matrix shows the effect of instead clamping player 2's move to the "average" action vector $\vec{a} = \{.33, .33, .33\}$, which amounts to replacing that player's move with the "illegal" move of fractionally taking each possible move ($\zeta_{\eta_2} = \vec{a}$).

probabilities of η's own moves. One such payoff function is the **Wonderful Life** Utility (WLU). The WLU for player η is parameterized by a pre-fixed **clamping parameter** CL_η chosen from among η's possible moves:

$$WLU_\eta \equiv G(\zeta) - G(\zeta_{\hat{\eta}}, CL_\eta) \ . \tag{7}$$

WLU is factored no matter what the choice of clamping parameter. Furthermore, while not matching the low opacity of AU, WLU usually has far better opacity than does a team game.

Figure 2 provides an example of clamping. As in that example, in many circumstances there is a particular choice of clamping parameter for player η that is a "null" move for that player, equivalent to removing that player from the system. (Hence the name of this payoff function — cf. the Frank Capra movie.) For such a clamping parameter assigning the associated WLU to η as its payoff function is closely related to the economics technique of "endogenizing a player's externalities" [17].

However it is usually the case that using WLU with a clamping parameter that is as close as possible to the expected move defining AU results in far lower opacity than does clamping to the null move. Such a WLU is roughly akin to a mean-field approximation to AU.[6] For example, in Fig. 2, if the probabilities of player 2 making each of its possible moves was $1/3$, then one would expect that a clamping parameter of $\vec{a}$ would be close to optimal. Accordingly, in practice use of such an alternative WLU derived as a "mean-field approximation" to AU almost always results in far better values of G than does the "endogenizing" WLU.

Intuitively, one can look at AU and WLU from the perspective of a human company, with G the "bottom line" of the company, the players η identified with the employees of that company, and the associated g_η given by the employees' performance-based compensation packages. For example, for a "factored company", each employee's compensation package

[6]Formally, our approximation is exact only if the expected value of G equals G evaluated at the expected joint move (both expectations being conditioned on given moves by all players other than η). In general though, for relatively smooth G, we would expect such a mean-field approximation to AU, to give good results, even if the approximation does not hold exactly.

contains incentives designed such that the better the bottom line of the corporation, the greater the employee's compensation. As an example, the CEO of a company wishing to have the payoff utilities of the employees be factored with G may give stock options to the employees. The net effect of this action is to ensure that what is good for the employee is also good for the company. In addition, if the compensation packages are "low opacity", the employees will have a relatively easy time discerning the relationship between their behavior and their compensation. In such a case the employees will both have the incentive to help the company and be able to determine how best to do so. Note that in practice, providing stock options is usually more effective in small companies than in large ones. This makes perfect sense in terms of the COIN formalism, since such options generally have lower opacity in small companies than they do in large companies, in which each employee has a hard time seeing how his/her moves affect the company's stock price.

3 The Bar Problem

Arthur's bar problem [1] can be viewed as a problem in designing COINs. Loosely speaking, in this problem at each time step each player η decides whether to attend a bar by predicting, based on its previous experience, whether the bar will be too crowded to be "rewarding" at that time, as quantified by a utility function G. The selfish nature of the players frustrates the global goal of maximizing G. This is because if most players think the attendance will be low (and therefore choose to attend), the attendance will actually be high, and vice-versa.

Here, we focus on the following six more general variants of the bar problem investigated in [33]: There are N players, each picking one out of seven moves every week. Each variant of the game is parameterized by $\ell \in \{1, 2, 3, 4, 5, 6\}$. In a given variant, each move of an agent corresponds to attending the bar on some particular subset of ℓ out of the seven nights of the current week (i.e., given ℓ, each possible move is an 'attendance profile' vertex of the 7-dimensional unit hypercube having ℓ 1's). In each week every player chooses a move. Then the associated payoffs for each player are communicated to that player, and the process is repeated. For simplicity, for each ℓ we chose the seven possible attendance profiles so that if the moves are selected randomly uniformly, the expected resultant attendance profile across all seven nights is also uniform. (For example, or $\ell = 2$, those profiles are (1, 1, 0, 0, 0, 0, 0), (0, 1, 1, 0, 0, 0, 0), etc.)

More formally, the world utility in any particular week is:

$$G(\zeta) \equiv \sum_{k=1}^{7} \phi(x_k(\zeta)) \ , \tag{8}$$

where $x_k(\zeta)$ is the total attendance on night k; ζ_η is η's move in that week; $\phi(y) \equiv y \exp{(-y/c)}$; and c is a real-valued parameter. Our choice of $\phi(.)$ means that when either too few or too many players attend some night in some week world utility G is low.

Since we wish to concentrate on the effects of the utilities rather than on the RL algorithms that use them, we use (very) simple RL algorithms.[7] We would expect that even marginally

[7] On the other hand, to use algorithms so patently deficient that they have never even been considered in the RL community — like the algorithms used in most of the bar problem literature — would seriously interfere with our ability to interpret our experiments.

more sophisticated RL algorithms would give better performance. In our algorithm each player η has a 7-dimensional vector giving its estimates of the utility it would receive for taking each possible move. At the beginning of each week, each η picks the night to attend randomly, using a Boltzmann distribution over the seven components of η's estimated utilities vector. For simplicity, temperature does not decay in time. However to reflect the fact that each player operates in a non-stationary environment, utility estimates are formed using exponentially aged data: in any week t, the estimate η makes for the utility for attending night i is a weighted average of all the utilities it has previously received when it attended that night, with the weights given by an exponential function of how long ago each such utility was. To form the players' initial training set, we had an initial period in which all moves by all players were chosen uniformly randomly, with no learning.

4 Experimental Results

We investigate three choices of clamping parameter: $\vec{0}$, $\vec{1} = (1,1,1,1,1,1,1)$, and the "average" move, $\vec{a} = \ell\frac{\vec{1}}{7}$, where as usual $\ell \in \{1,2,3,4,5,6\}$, depending on the problem. (To keep the "congestion" level of the different problems close to one another, for ℓ going from 1 to 6, $c = \{3,6,8,10,12,15\}$ respectively.) The associated WLU's are distinguished with a superscript. In each of the experiments reported here all players have the same utility function, so from now on we drop the player subscript from the payoff utilities. Writing them out, the three WLU functions are:

$$
\begin{aligned}
WLU^{\vec{0}}(\zeta) &\equiv G(\zeta) - G(\zeta_{\hat{\eta}}, \vec{0}) \\
&= \phi(x_{d_\eta}(\zeta)) \;-\; \phi(x_{d_\eta}(\zeta) - 1) \\
WLU^{\vec{1}}(\zeta) &\equiv G(\zeta) - G(\zeta_{\hat{\eta}}, \vec{1}) \\
&= \sum_{d \neq d_\eta}^{7} \left(\phi(x_d(\zeta)) \;-\; \phi(x_d(\zeta) + 1)\right) \\
WLU^{\vec{a}}(\zeta) &\equiv G(\zeta) - G(\zeta_{\hat{\eta}}, \vec{a}) \\
&= \sum_{d=1}^{7} \phi(x_d(\zeta)) \;-\; \sum_{d \neq d_\eta}^{7} \phi\left(x_d(\zeta) + \frac{\ell}{7}\right) \;-\; \phi\left(x_{d_\eta}(\zeta) - 1 + \frac{\ell}{7}\right)
\end{aligned}
$$

where d_η is the night picked by η.

Based on the analysis presented above, $WLU^{\vec{a}}$ is the WLU that we would expect to be optimal if the probability of each move by any particular player were $1/7$, since in those circumstances it is the mean-field approximation to AU. In practice of course, it is not true that each player has a uniform distribution over its possible moves. Here, to avoid addressing the self-consistency issue associated with evaluating AU, we choose to handicap the algorithms based on the predictions of our mathematics by making the (clearly very crude) approximation that each player's probability distribution is indeed uniform, in addition to making the mean-field approximation.

In that it subtracts from the actual value of G what G would have been if player η had never existed, $WLU^{\vec{0}}$ is akin to the standard economics technique of "endogenizing player η's externalities". Note that to evaluate it for player η one only needs to know the total attendance on the night(s) attended by η. In contrast, G (team game utility) and $WLU^{\vec{a}}$ require centralized communication concerning all 7 nights, and $WLU^{\vec{1}}$ requires communication concerning 6 nights.

In the first experiment each player had to select one night to attend the bar ($\ell = 1$ and the vertex is the identity matrix.) In this case $CL_\eta = \vec{0}$ is equivalent to the player "staying at home". On the other hand, $CL_\eta = \vec{1}$ corresponds to the player attending every night. Finally, $CL_\eta = \vec{a} = \frac{\vec{1}}{7}$ is equivalent to the players attending partially on all nights in proportions equivalent to the overall attendance profile of all players across the initial training period. (Note that none of these "moves" are actually available to the players. Rather these fictional moves are used to compute the players' payoff utilities, as described in Section 2.)

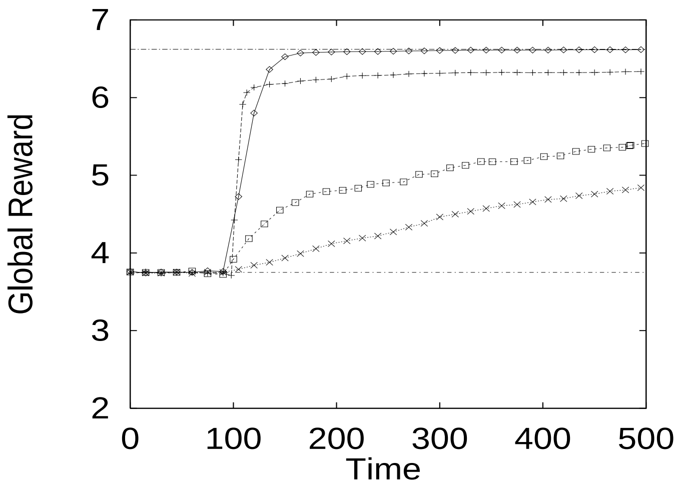

Figure 3: Effect of payoff utility functions on system performance; $\ell = 1$; ($WLU^{\vec{a}}$ is $\diamond$; $WLU^{\vec{0}}$ is $+$; $WLU^{\vec{1}}$ is $\square$; G is $\times$)

Figure 3 graphs world utility against time, averaged over 100 runs, for 60 players and $c = 3$. (Throughout this paper, error bars are too small to depict.) The two straight lines correspond to the optimal performance, and the "baseline" performance given by uniform occupancies across all nights. Systems using $WL_{\vec{a}}$ and $WL_{\vec{0}}$ rapidly converged to optimal and to quite good performance, respectively. This indicates that for the bar problem the "mild assumptions" mentioned above hold, and that the approximations in the derivation of the optimal clamping parameter are valid.

Figure 4 shows the normalized world utility obtained for the different payoff utilities as a function of ℓ (i.e., when players attend the bar on ℓ nights in one week). Temperatures varied between .01 and .02 for the three WL utilities, and between .1 and .2 for the G utility, which provided the respective best performances for each. $WLU^{\vec{a}}$ performs well for all problems. $WLU^{\vec{1}}$ on the other hand performs poorly when players only attend on a few nights, but reaches the performance of $WLU^{\vec{a}}$ when players need to select six nights, a situation where the two clamping vectors are very similar ($\vec{1}$ and $\frac{\vec{6}}{7}$, respectively). $WLU^{\vec{0}}$ shows a slight drop in performance when the number of nights to attend increases. The fact that it always performs worse than does $WLU^{\vec{a}}$ illustrates the shortcoming of conventional economics techniques which do not consider the kind of signal-to-noise issues that drive opacity.

G shows an even more pronounced drop with the number of nights to attend than does

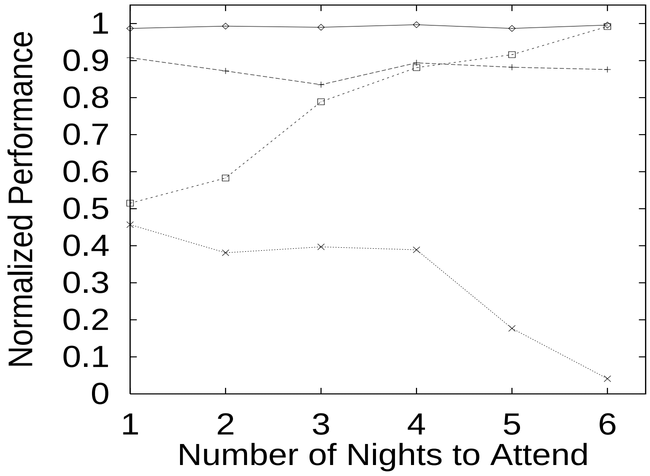

Figure 4: Behavior of payoff utility functions with respect to number of nights to attend ($\ell = 1$). ($WLU^{\vec{a}}$ is $\diamond$; $WLU^{\vec{0}}$ is $+$; $WLU^{\vec{1}}$ is $\square$; G is $\times$)

$WLU^{\vec{0}}$. Furthermore, in agreement with our previous results [33], for all problems the poor signal-to-noise when using the payoff function G results in poor values of world utility, despite that payoff function's being factored. Taken together, these results confirm our theoretical prediction of what payoff utility converges fastest to the world utility maximum.

Figure 5 shows how $t = 500$ performance scales with N for each of the utility functions. For comparison purposes the performance is normalized — for each utility U we plot $\frac{U - G_{base}}{G_{opt} - G_{base}}$, where G_{opt} and G_{base} are the optimal performance and a canonical baseline performance given by uniform attendance across all nights, respectively. Systems using team game utility (G) perform adequately when N is low. As N increases however, it becomes increasingly difficult for the players to extract the information they need from G. Because of their lower opacity, systems using the different WLUs overcome this signal-to-noise problem to a great extent. Because the WLU is based on the *difference* between the actual state and the state where one player's state is clamped, they are much less affected by the total number of players. However, note that vector to which players are clamped significantly affects the scaling properties, showing that even among difference utilities, optimizing for opacity provides a significant gain.

We also studied the sensitivity of performance to the internal parameters of the learning algorithms. Figure 6 presents experiments with $\ell = 1$ for a set of different temperatures in the RL algorithms. (The two straight lines correspond to the optimal performance, and the "baseline" performance given by uniform occupancies across all nights.) $WLU^{\vec{a}}$ is fairly insensitive to the temperature, until it gets so high that players' moves are chosen almost randomly. $WLU^{\vec{0}}$ depends more than $WLU^{\vec{a}}$ does on having sufficient exploration and therefore has a narrower range of good temperatures. Both $WLU^{\vec{1}}$ and G have more serious opacity problems, and therefore have shallower and thinner performance graphs.

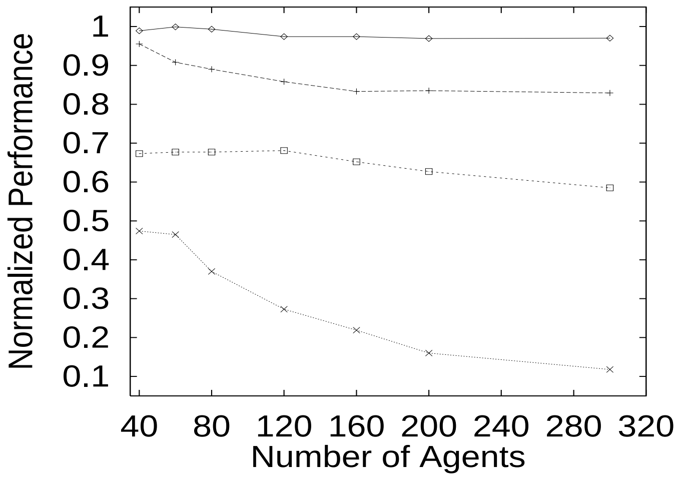

Figure 5: Scaling properties of the payoff utility functions; $\ell = 1$; $c = \{2, 3, 4, 6, 8, 10, 15\}$ for $N = \{40, 60, 80, 120, 160, 200, 300\}$, respectively. ($WLU^{\vec{a}}$ is $\diamond$; $WLU^{\vec{0}}$ is $+$; $WLU^{\vec{1}}$ is $\square$; G is $\times$)

5 Conclusion

In this article we considered how to design large multi-agent systems to meet a pre-specified goal when each agent in the system uses reinforcement learning to choose its actions. We cast this problem as how to initialize/update the individual agents' payoff utility functions so that their collective behavior optimizes a pre-specified world utility function. The mathematics of COINs is specifically concerned with this problem. In previous experiments we showed that systems based on that math far outperformed conventional "team game" systems, in which each agent has the world utility as its private utility function. Moreover, the gain in performance grows with the size of the system, typically reaching orders of magnitude for systems that consist of hundred of agents.

In those previous experiments the COIN-based private utilities had a free parameter, which we arbitrarily set to 0. However as synopsized in this paper, it turns out that a series of approximations allows one to derive an optimal value for that parameter. Here we have repeated some of our previous computer experiments, only using this new value for the parameter. These experiments confirm that with this new value the system converges to significantly superior world utility values, with less sensitivity to the parameters of the agents' RL algorithms. This makes even stronger the arguments for using a COIN-based system rather than a team-game system. Future work involves improving the approximations needed to calculate the optimal private utility parameter value. In particular, given that that value varies in time, we intend to investigate having it be calculated in an on-line manner.

References

[1] W. B. Arthur. Complexity in economic theory: Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, May 1994.

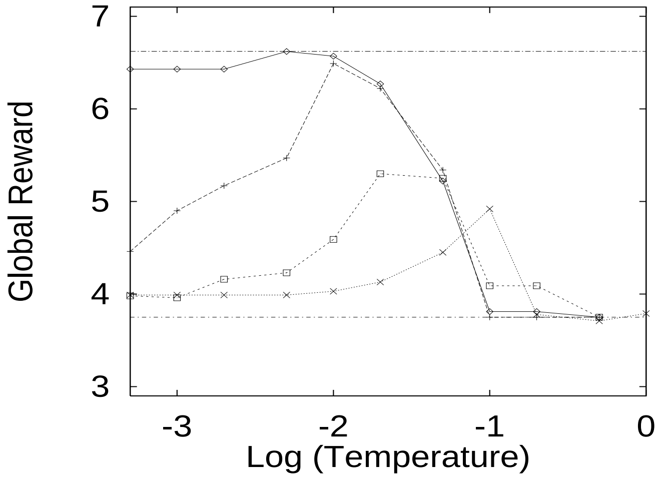

Figure 6: Sensitivity of payoff utility functions to internal parameters; $\ell = 1; c = 3$. ($WLU^{\vec{a}}$ is $\diamond$; $WLU^{\vec{0}}$ is $+$; $WLU^{\vec{1}}$ is $\Box$; G is $\times$)

[2] C. Boutilier. Multiagent systems: Challenges and opportunities for decision theoretic planning. *AI Magazine*, 20:35–43, winter 1999.

[3] C. Boutilier, Y. Shoham, and M. P. Wellman. Editorial: Economic principles of multi-agent systems. *Artificial Intelligence Journal*, 94:1–6, 1997.

[4] J. M. Bradshaw, editor. *Software Agents.* MIT Press, 1997.

[5] G. Caldarelli, M. Marsili, and Y. C. Zhang. A prototype model of stock exchange. *Europhysics Letters*, 40:479–484, 1997.

[6] D. Challet and Y. C. Zhang. On the minority game: Analytical and numerical studies. *Physica A*, 256:514, 1998.

[7] C. Claus and C. Boutilier. The dynamics of reinforcement learning cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, Madison, WI, June 1998.

[8] R. H. Crites and A. G. Barto. Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems - 8*, pages 1017–1023. MIT Press, 1996.

[9] D. Fudenberg and J. Tirole. *Game Theory.* MIT Press, Cambridge, MA, 1991.

[10] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250, June 1998.

[11] B. A. Huberman and T. Hogg. The behavior of computational ecologies. In *The Ecology of Computation*, pages 77–115. North-Holland, 1988.

[12] N. R. Jennings, K. Sycara, and M. Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1:7–38, 1998.

[13] N. F. Johnson, S. Jarvis, R. Jonson, P. Cheung, Y. R. Kwong, and P. M. Hui. Volatility and agent adaptability in a self-organizing market. preprint cond-mat/9802177, February 1998.

[14] L. P. Kaelbing, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[15] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 157–163, 1994.

[16] D. Monderer and L. S. Sharpley. Potential games. *Games and Economic Behavior*, 14:124–143, 1996.

[17] W. Nicholson. *Microeconomic Theory*. The Dryden Press, seventh edition, 1998.

[18] T. Sandholm and R. Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37:147–166, 1995.

[19] T. Sandholm, K. Larson, M. Anderson, O. Shehory, and F. Tohme. Anytime coalition structure generation with worst case guarantees. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 46–53, 1998.

[20] S. Sen. *Multi-Agent Learning: Papers from the 1997 AAAI Workshop (Technical Report WS-97-03*. AAAI Press, Menlo Park, CA, 1997.

[21] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[22] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[23] K. Sycara. Multiagent systems. *AI Magazine*, 19(2):79–92, 1998.

[24] K. Tumer and D. H. Wolpert. Collective intelligence and Braess' paradox. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 104–109, Austin, TX, 2000.

[25] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3/4):279–292, 1992.

[26] M. P. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. In *Journal of Artificial Intelligence Research*, 1993.

[27] D. H. Wolpert. Bounded-rationality game theory. pre-print, 2001.

[28] D. H. Wolpert. The mathematics of collective intelligence. pre-print, 2001.

[29] D. H. Wolpert, E. Bandari, and K. Tumer. Improving simulated annealing by recasting it as a non-cooperative game. *Nature*, 2001. submitted.

[30] D. H. Wolpert and K. Tumer. An Introduction to Collective Intelligence. Technical Report NASA-ARC-IC-99-63, NASA Ames Research Center, 1999. URL:`http://ic.arc.nasa.gov/ic/projects/coin_pubs.html`. To appear in Handbook of Agent Technology, Ed. J. M. Bradshaw, AAAI/MIT Press.

[31] D. H. Wolpert, K. Tumer, and J. Frank. Using collective intelligence to route internet traffic. In *Advances in Neural Information Processing Systems - 11*, pages 952–958. MIT Press, 1999.

[32] D. H. Wolpert, K. Wheeler, and K. Tumer. General principles of learning-based multi-agent systems. In *Proceedings of the Third International Conference of Autonomous Agents*, pages 77–83, 1999.

[33] D. H. Wolpert, K. Wheeler, and K. Tumer. Collective intelligence for control of distributed dynamical systems. *Europhysics Letters*, 49(6), March 2000.

[34] Y. C. Zhang. Modeling market mechanism with evolutionary games. *Europhysics Letters*, March/April 1998.