Learning and Decision-Making for Intention Reconciliation

Learning and Decision-Making for Intention Reconciliation

Sanmay Das
MIT Center for Biological and
Computational Learning
E25-201, 45 Carleton St.
Cambridge, MA 02142
sanmay@ai.mit.edu

Barbara Grosz
Division of Engineering and
Applied Sciences, Harvard
University
33 Oxford Street
Cambridge, MA 02138
grosz@eecs.harvard.edu

Avi Pfeffer
Division of Engineering and
Applied Sciences, Harvard
University
33 Oxford Street
Cambridge, MA 02138
avi@eecs.harvard.edu

ABSTRACT

Rational, autonomous agents must be able to revise their commitments in the light of new opportunities. They must decide when to default on commitments to the group in order to commit to potentially more valuable outside offers. The SPIRE experimental system allows the study of intention reconciliation in team contexts. This paper presents a new framework for SPIRE that allows for mathematical specification and provides a basis for the study of learning. Analysis shows that a reactive policy can be expected to perform as well as more complex policies that look ahead. We present an algorithm for learning when to default on group commitments based solely on observed values of group-related tasks and discuss the applicability of this algorithm in settings where multiple agents may be learning.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Design, Economics, Performance, Theory

Keywords

Evolution, adaptation and learning; group and organizational dynamics

1. INTRODUCTION

Sullivan et al. [12], citing Grosz and Kraus [6] and Bratman [2], note that rational agents cannot adopt conflicting intentions. If an agent has adopted an intention to perform some task that is part of a group activity, thereby committing to that task, it cannot also adopt an intention to perform some other activity that conflicts with that task. Because the agent cannot commit to both the group-related activity and the other activity, it must decide whether to maintain its intention to perform the group-related task or renege on its commitment to the group task and adopt an intention to perform the other activity. It must *reconcile* these intentions.

The SPIRE (SharedPlans Intention Reconciliation Experiments) simulation system was designed to allow the study of the effects of different environmental parameters and agent decision-making procedures in a collaborative setting [12, 5, 13, 7]. The system enables investigation of the problem of when it is appropriate for an agent to default on a commitment to a *group-related task* in order to commit to a potentially more valuable *outside offer*.

One key element of the SPIRE framework is a notion of reputation for each agent. An agent's reputation is hurt whenever it defaults on a commitment to a group-related task, but previous "bad behavior" is gradually forgiven over time. The average utility of group-related tasks an agent receives is a function of its reputation. In the model used for this work (following Grosz et al. [7]), the values of tasks each agent receives are also dependent on the reputation of the group as a whole. Hence the behavior of each agent affects not only the utility it can expect to receive in the future from group-related activities, but also the utility each and every member of the group can expect to receive in the future from group-related activities.

Within this framework, simulations are performed in which agents interact repeatedly with the environment by choosing whether or not to default on commitments to group-related tasks. The framework allows for the simulation of heterogeneous communities of agents that have different decision making procedures. Agent performance is typically measured by average income (or reward) received over the course of the simulation.

This paper presents a new model for the study of intention reconciliation. The framework builds on the previous model of agent-group interaction in the SPIRE system, but simplifies the model considerably and provides a framework for the study of learning. The number of parameters in the system is significantly reduced, making the model more conducive to analytical treatment while isolating the most important aspects of agent-group interaction.

In this new model, the fully-informed decision-making strategy used by Grosz et al. [7] reduces to a simple strategy agents can use to decide whether or not to default on a group commitment without using any information about the state of the world. This strategy involves maintaining a *cut-*

AAMAS'02, July 15-19, 2002, Bologna, Italy.
Copyright 2002 ACM 1-58113-480-0/02/0007 ...$5.00.

off value and defaulting whenever the utility of the outside offer exceeds the utility of the group-related task by more than that value. We present empirical results for communities of homogeneous and heterogeneous agents using such cutoff functions. The results for homogeneous communities show that there is a cutoff that optimizes mean individual utility received by agents and that this is a global maximum.

It is unrealistic for designers of real-world multi-agent systems to evaluate many different decision-making procedures through extensive simulations to decide on the optimal procedure. Further, the composition of a group of agents or the exact parameters and model of the environment may be unknown at the time agents are designed. It is necessary for agents to adapt successfully to different environments. We define the learning problem in the new framework and present an algorithm that learns the cost of defaulting. We analyze the performance of this algorithm in situations where single and multiple agents are learning.

2. RELATED WORK

Several recent approaches to modeling the intention reconciliation problem use a market-oriented or contract-based framework. Sandholm and Lesser [9] introduce the concept of *leveled commitment contracts* in which agents pay the rest of the group a predetermined penalty for defaulting on a group-related task. Sen and Biswas [10] introduce a setup in which there is no direct market-mediation, but instead agents choose which others to work with based on previous experiences. Teague and Sonenberg [14] investigate the effects of different ways of imposing penalties for defaulting in the context of a target-capturing game.

The SPIRE project, to which the work presented in this paper belongs, provides a general framework from which to analyze and model the intention reconciliation problem. SPIRE examines the interactions between an agent and the entire group to which it belongs, rather than interactions between individual agents. Also, SPIRE examines the behavior of agents that are committed to working in a particular group. They do not have the option of leaving the group altogether. The best analogy for the SPIRE framework is that of a company or firm that gets jobs from the market and then contracts tasks required for the jobs to its agents.

Previous work using SPIRE considered three different ways of imposing penalties for defaulting:

- A penalty for reneging on group-related activities that represents the cost to the group, which is shared across the entire group.
- The allocation of a percentage of all tasks such that agents who default less get more valuable tasks.
- A measure of self-imposed reluctance to default[1].

Our research also investigates learning in multi-agent settings. In the single-agent setting, reinforcement learning addresses the question of how an autonomous agent that senses and acts in an environment can learn to choose actions that maximize its utility. The agent typically learns a *policy*, a mapping from states to actions that specifies the action the agent should take in any given state. Most work in reinforcement learning uses the *Markov Decision Process* (MDP) framework. The Markov property states that the probability of a transition from one state to another depends solely on the current state and action, not on the history of actions an agent has taken or states it has visited [8].

Stochastic games, also known as *multi-agent MDPs*, are a natural extension of the MDP framework to multi-agent domains. Stochastic games (SGs), first introduced by Shapley [11], have been studied in the economics and game theory communities. However, the initial work of Shapley, and further work such as that of Vrieze [15], assumes that the agent knows the transition and reward models, which is not the case in the environments studied here. Recently, reinforcement learning techniques have also been applied to SGs [3, 1]. Learning in multi-agent environments is difficult because of their potential non-stationarity. If an agent is learning in an environment where its payoffs are affected by other agents, the environmental parameters are significantly affected by the behavior of the other agents. This is not a problem when all the other agents are following static or fixed policies. However, if the other agents are also learning about the environment and modifying their strategies over time, none of the standard reinforcement learning algorithms are guaranteed to converge [1].

3. THE SPIRE MODEL

3.1 The Initial Model

In SPIRE, a group of agents works together on a set of group tasks. An agent performs a group-related activity or group-related task (sometimes referred to simply as tasks) as part of a larger group task. Outside offers (referred to as offers) are generated a fixed percentage of the time and are offered to a random agent. When an agent receives an offer for a time slot in which a task is already scheduled, it must decide whether to renege on the group-related task to accept the outside offer or not. Another agent may be able to replace an agent that defaults on a commitment to a group-related activity[2]. Each default incurs a baseline cost that is shared across the whole group. In addition, a portion of tasks are distributed to agents based on their ranking by reputation, and thus agents with better reputations get higher valued tasks.

Because of the many parameters and different effects involved in the original SPIRE model, it is difficult to reason about the causes of various effects. Further, recent work by Teague and Sonenberg [14] suggests that different methods of imposing penalties like the two mentioned above, as well as the self-imposed reluctance of agents to default (as discussed in the "brownie point" model of Glass and Grosz [5]) lead to similar defaulting dynamics in the group.

3.2 A New Model for Reputation, Group-Related Tasks and Outside Offers

3.2.1 The Market-Oriented Motivation

The framework we develop models a group (for example a company or firm) which receives a certain number of tasks to be performed at each time period from some outside contracting agency. The group distributes these tasks among its own workers, who cannot refuse a task assignment at the time when tasks are assigned. However, the workers may

[1] This "good-guy" effect, modeled using "brownie points" is described in detail by Glass [4] and Glass and Grosz [5].

[2] For more details see Sullivan et al. [12] and Grosz et al. [7]

receive outside offers that they can then take. If an outside offer conflicts with a group-related task that the agent is supposed to be performing, the agent may renege on the group-related task.

The market will be unwilling to pay as much to a group with a poor reputation as it would pay to a group with a good reputation. A group can acquire a poor reputation because many of its employees renege on their group commitments on a regular basis, leading to tasks not being completed. In the same way, a group has the ability to improve its reputation and get paid more for completing an equivalent set of tasks. So a group has the ability to change the amount of income it makes from the tasks it receives.

Suppose the group receives an income I from a particular task performed by a member-agent. The group is unlikely to give the agent the entire sum I. Instead it will give the agent some proportion of I, keeping the rest for operational expenses and overhead. The amount the group keeps is determined by the individual reputation of the agent within the group. The group will be willing to pay less to an individual with a poor reputation because of the perceived risk. The upper bound on what the agent can make from the task is, of course, I. It is important to note that, in this framework, the overhead the company keeps can be thought of as a cost associated with agent defaults. It is not intended to be a profit margin that the company management tries to maximize.

3.2.2 Group-Related Tasks and Outside Offers

Group task values are drawn from a truncated normal distribution, and reputation penalties are applied to all task values. Initial values are drawn from a normal distribution with a specified mean $\mu = M_0$ and standard deviation σ. Whenever the value generated is below 0, the task value is set to 0. A value drawn from this distribution represents the average "market-value" of a task. Depending on the reputation of the group, the group could be offered more or less than this value to perform the task. Let group reputation be $\mathbf{GR}$, and let α be a parameter associated with $\mathbf{GR}$. Then, for performing a task with average market value M, the group will get $M + \alpha\mathbf{GR}$. The value of performing this task to an agent a is given by $M + \alpha\mathbf{GR} + \beta\mathbf{IR}_a$, where $\mathbf{IR}_a$ is a measure of the individual reputation of agent a and β is an associated scaling parameter.

Outside offers are also drawn from a normal distribution. The distribution of outside offer values are typically set to have a lower mean and higher standard deviation than the distribution of group task values. An outside offer is generated for each agent at each time slot.

Reputation changes are effectively instantaneous, since new task and offer values for the next time slot take into account the changed reputations. γ is defined as a parameter in the range $[0, 1]$ that allows "forgetting" of previous defaults, so that reputation can improve over time.

Group reputation $\mathbf{GR}$ is defined as a value in the range $[-0.5, 0.5]$. The evolution of $\mathbf{GR}$ from one time period t to the next, $t + 1$ is governed by the equation:

$$\mathbf{GR}^{t+1} = \gamma\mathbf{GR}^t + (1-\gamma)(-0.5\frac{\textit{num-defaults}^t}{\textit{num-agents}})$$

where $\textit{num-defaults}^t$ represents the number of defaults that occurred in the entire group at time t. Thus, group reputation will be stable at around 0 if approximately half the members of the group default all the time, and will asymptotically approach 0.5 if none of the members ever default and -0.5 if all the members always default. As a result, $\mathbf{GR}$ is robust to many different environments in terms of outside offer means and standard deviations.

The individual reputation of an agent a at time t, $\mathbf{IR}_a^t$ is defined in the range $[-1, 0]$ so that it always serves as a negative factor, and an agent cannot make more than its group was given for a particular task. $\mathbf{IR}$ evolves as follows:

$$\mathbf{IR}_a^{t+1} = \gamma\mathbf{IR}_a^t + (1-\gamma)(-\textit{default}(a,t))$$

where $\textit{default}(a,t)$ is an indicator function that is 1 if agent a defaulted at time t and 0 otherwise. This function is again symmetric and linear.

3.2.3 Changes from the Original System

Our new model differs significantly from the model presented by Grosz et al. [7] in several respects while preserving the basic dynamics of agent-group interactions. The major differences in the models include:

- Group task and outside offer values are drawn from normal distributions truncated at 0, rather than from uniform distributions.

- In the original SPIRE framework, outside offers were only generated a percentage of the time as controlled by an "outside offer percentage" parameter. A similar effect is achieved in our model. Since it is never advantageous for an agent to default on a group-related task unless the competing outside offer is more valuable, the agent will only consider some percentage of the outside offers. For any choice of means and standard deviations for task values and offer values, a certain percentage of outside offers will be more valuable than the group-related tasks.

- When only a small proportion of group-related tasks are assigned based on reputation, observations of individual utility received tend to be damped by the randomly distributed tasks. To remove the damping effect, reputation penalties are applied to all tasks rather than to a fixed proportion of tasks in the new model.

- The week-based framework in which there were 40 time slots per week has been eliminated. Reputations now change after every time period rather than at the end of a "week" consisting of 40 time periods.

- The "task-density" parameter of the original framework, which controlled the percentage of time slots in which an agent would have a group-related task to perform, is unnecessary in the new model. If the task-value to an agent is 0, which corresponds to a situation where the agent is not given a task, the agent is free to take an outside offer without having its reputation suffer. As a result the task-density parameter arises naturally[3].

[3] The manner in which the task-density parameter arises can lead to situations in which a task with value to the group has no value for a particular agent because of the $\beta\mathbf{IR}_a$ factor. In these cases, the group refuses to assign the task to the agent due to the risk.

- The new model uses a linear reputation function. Nonlinear functions are useful in certain environments like 40 time-slot based weeks, because the change of reputation does not occur till the end of the week, and an agent has the opportunity to default multiple times in one week. If the reputation function were linear, an agent that defaulted once would almost always default all the time, leading to extreme behaviors in most cases. In an immediate-update scenario, the forgetting factor is strong enough to allow different agents to have IRs in different ranges without seeing extreme behavior like always defaulting or never defaulting.

Our model is simpler and more streamlined than the original SPIRE model, with only seven parameters: the means and variances of the group task and outside offer values, the scaling parameters α and β, and the forgiveness factor γ. Despite the simplicity, our model captures key features of the intention reconciliation problem, such as the effect of defaulting on both the individual and the group, and the need to consider the long-term effects of a short-term decision to default.

4. STATIC DECISION-MAKING POLICIES

4.1 Decision-Making Using One Step Lookahead and Cutoff Functions

In previous work on SPIRE, Grosz et al. [7] use a decision-making function that assumes full knowledge of both the state (including the individual and group reputations) and the equations which determine the state transitions (the *transition model*). This section defines a similar decision-making function **DF** that is essentially a direct translation of the utility-based decision-making function used in previous work to the new model. This informed decision-making function is provably equivalent to a decision-making function that does not need any information about the transition model or the state, but simply determines whether to default or not based on whether the difference in the outside offer value and the task value to the agent is greater than a specified cutoff. This section presents empirical results for homogeneous and heterogeneous communities of agents using such "cutoff functions."

4.1.1 *The Decision Making Function* $\mathbf{DF}(\delta)$ *and the Cutoff Function* $\mathbf{CF}(v)$

The decision making function $\mathbf{DF}(\delta)$ is based on one-step lookahead. The parameter δ specifies the weight an agent attaches to its estimate of future expected income. We stipulate that agents are not aware of the finite horizon when their interaction with the group will come to an end, because finite horizons can lead to significant changes in defaulting behavior, especially towards the end of a simulation [5]. Therefore, agents always make the assumption that their interaction with the group will continue for an infinite number of time-steps.

$\mathbf{DF}(\delta)$ computes future expected income (FEI) in the cases where an agent defaults and those in which it does not default. FEI does not take into consideration future outside offers, it only accounts for future group tasks. The expected income for the next week is computed for the above two cases. As discussed above, the agent knows the mean of the distribution that initial task-values are drawn from and the parameters α and β, along with $\mathbf{IR}_a^t$ and $\mathbf{GR}^t$. This income is then multiplied by a weighting factor based on δ. In the infinite horizon case we consider, this weighting factor is the sum of the infinite geometric progression $\delta, \delta^2, \delta^3, \ldots$, which is $\frac{\delta}{1-\delta}$. If the weighted difference between the estimated FEI given the agent does not default and the weighted FEI given that the agent does default is greater than the gain in CI that can be obtained by defaulting, the agent does not default, otherwise it does.

The cutoff function $\mathbf{CF}(v)$ is defined such that an agent using $\mathbf{CF}(v)$ as its decision function chooses to default whenever the difference between the outside offer value and the task value is greater than v.

4.1.2 *The Function* $\mathbf{DF}(\delta)$ *is Equivalent to a Cutoff Function*

Suppose the task value for time period t is T and the offer value is O. M_0 is the mean from which task values are drawn before factoring in reputations, $\mathbf{GR}^t$ is the group reputation at time t, $\mathbf{IR}_a^t$ is the individual reputation of agent a at time t, δ is the parameter used by agents using **DF** to discount future expected income, and γ is the "forgetting" parameter in the evolution of **IR** and **GR**.

Define $\mathbf{IR}_a^{t+1}(\mathit{nodef})$ to be the reputation of agent a if it does not default at time t and $\mathbf{IR}_a^{t+1}(\mathit{def})$ to be the reputation if it does default at time t. Let FEI-NODEF be the weighted expected future income as defined by the function **DF** if the agent does not default at time t and let FEI-DEF be the weighted expected future income as defined by DF if the agent defaults at time t. Then:

$$\text{FEI-NODEF} = \frac{\delta}{1-\delta}(M_0 + \beta\mathbf{IR}_a^{t+1}(\mathit{nodef}) + \alpha\mathbf{GR}^t)$$

$$\text{FEI-DEF} = \frac{\delta}{1-\delta}(M_0 + \beta\mathbf{IR}_a^{t+1}(\mathit{def}) + \alpha\mathbf{GR}^t)$$

Therefore, the difference in FEI is given by

$$\text{FEI-DIFF} = \frac{\delta}{1-\delta}(\beta\mathbf{IR}_a^{t+1}(\mathit{nodef}) - \beta\mathbf{IR}_a^{t+1}(\mathit{def}))$$

Now, $\mathbf{IR}_a^{t+1}(\mathit{nodef}) = \gamma\mathbf{IR}_a^t$ and $\mathbf{IR}_a^{t+1}(\mathit{def}) = \gamma\mathbf{IR}_a^t + \gamma - 1$. Therefore:

$$\text{FEI-DIFF} = \frac{\delta\beta}{1-\delta}[(\gamma\mathbf{IR}_a^t - \gamma\mathbf{IR}_a^t) + (1-\gamma)]$$

$$\Rightarrow \text{FEI-DIFF} = \frac{\beta(1-\gamma)\delta}{1-\delta}$$

Therefore, if $O - T > \frac{\beta(1-\gamma)\delta}{1-\delta}$, the agent will choose to default. The function $\mathbf{DF}(\delta)$ is equivalent to the function $\mathbf{CF}(\frac{\beta(1-\gamma)\delta}{1-\delta})$.

Note that the exact form of the result presented here does not always hold, because if α and β are sufficiently high and $\mathbf{IR}_a$ and $\mathbf{GR}$ sufficiently low, the values of FEI-NODEF and FEI-DEF may go below 0 in the above analysis, while the agent will always assume that the minimum income it can get from group-related tasks is 0, not negative. However, this case does not arise for any of the parameter settings discussed in this paper.

This result has multiple implications. First, an agent using the best simple cutoff rule will perform as well as any agent that performs one step lookahead. Second, a cutoff

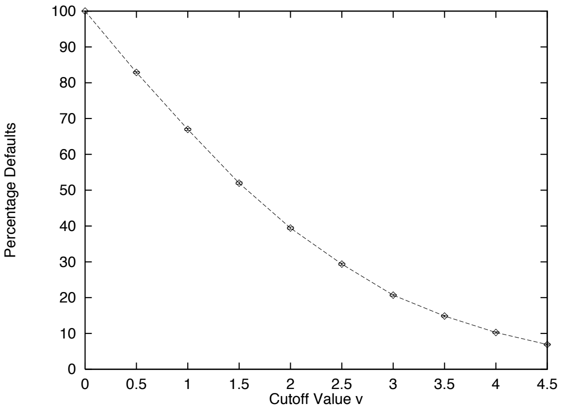

Figure 1: Percentage of outside offers accepted by agents, leading to defaults on group-related tasks, as a function of cutoff value

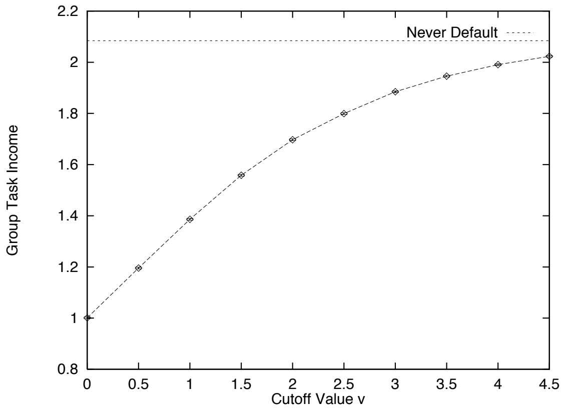

Figure 2: Group task incomes as a function of cutoff value

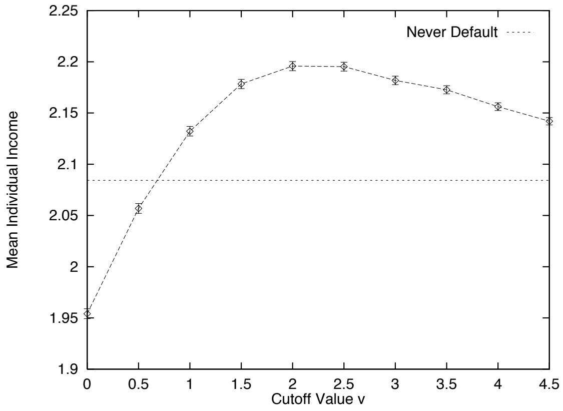

Figure 3: Mean individual incomes as a function of cutoff value

agent does not need to know **GR** and **IR**, so a good policy can be implemented even when **GR** and **IR** are not known. Third, since the optimal cutoff does not depend on **IR**, a rational cutoff agent will use the same cutoff no matter what its individual reputation. As a result, we avoid the problem of *dropouts*, in which an agent's reputation becomes so poor that it considers itself unredeemable and always takes the outside offer.

4.2 Experimental Methodology

The next two sections present experimental results. Experiments reported in this section all use the same basic parameters. Unless otherwise specified, $\alpha = 2$; $\beta = 0.5$; the mean of the distribution from which group-related task values are drawn, $M_0 = 1$; the standard deviation of this distribution, $\sigma_{\text{gt}} = 1$; the mean of the distribution from which offer values are drawn, $\mu_{\text{oo}} = 0$; the standard deviation $\sigma_{\text{oo}} = 3$. The heavy weight α on **GR** replaces to some extent the notion of group costs in the original SPIRE model. Outside offers that are more valuable than the group-related activity assigned to an agent at the same time do not arise frequently, but they have the potential to afford the agent considerable utility. Experiments were run for 1000 time periods, and the first 100 time periods were ignored in tabulating average incomes or rewards to allow the group and individual reputations to stabilize. Because of the static decision-making functions, allowing only 100 periods for this stabilization and gathering data from 1000 weeks leads to results that are the same as those obtained by running for more time periods. Error bars represent 95% confidence intervals.

4.3 Homogeneous Communities of Agents Using Cutoff Functions

These experiments investigate the behavior of communities of agents in which all the agents use the same value of v for their cutoff function $\mathbf{CF}(v)$. Agents using a lower value of v will be less "socially responsible" in the sense used by Glass [4], that is, they will tend to default more on their group commitments, because it requires a lower incentive for them to be willing to renege on the group.

Figure 1 shows this effect clearly. The percentage of defaults decreases as the cutoff value v increases, asymptotically approaching 0. Group task income is defined as the income the group gets from performing group tasks only. The value of group task income is closely related to the percentage of defaults, and one can see from Figure 2 that it increases as the percentage of defaults decreases, asymptotically reaching the theoretical optimum, which occurs when none of the agents ever default on group commitments.

The most interesting result is shown in Figure 3, which shows the mean individual income of group members as a function of cutoff values. Individual income is defined as the sum of the incomes an agent receives from group-related tasks and outside offers (keeping in mind that the agent receives less income from completing a group task than the group does). The optimal level of defaulting lies in between never defaulting and always defaulting. This not only corroborates the results of Grosz et al. [7] and Glass and Grosz [5] that an intermediate level of "social consciousness" is often optimal, it also shows that these results are robust across different models of agent-group interaction, and that they are a basic feature of domains like SPIRE in which there are environmental incentives not to default. In the agent-agent model of interaction of Teague and Sonenberg [14], self-imposed reluctance to default (which is present in

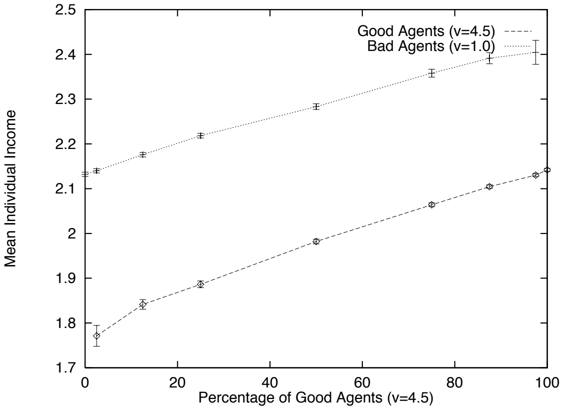

Figure 4: Mean individual incomes of good and bad agents that form a mixed community as a function of the percentage of agents that are good

the brownie points model) leads to results similar to those produced by the presence of environmental incentives not to default. The results presented in this section extend this finding to the SPIRE model of agent-group interaction.

4.4 Heterogeneous Communities of Agents Using Cutoff Functions

This section examines the behavior of agents in mixed communities. For the sake of simplicity and establishing some basic results, the experiments focus on the case where a group consists of two types of agents, and the percentage of each type changes. We use representatives of "good" agents, those who are socially responsible and do not default much, and "bad" agents who default more frequently. For the purposes of comparison, two cutoff values are used such that homogeneous communities of agents using these two cutoffs perform at about the same level in terms of mean individual income. The two cutoff values are $v = 1.0$ (bad agents) and $v = 4.5$ (good agents). A homogeneous community of agents using $v = 4.5$ performed somewhat better than a homogeneous community of agents using $v = 1.0$.

One can see from Figure 4 that the bad agents outperformed the good agents for any particular percentage of good agents. Therefore, it is individually optimal for any single agent to default frequently in any particular mix of good and bad agents. In fact, a single agent that always defaults (i.e. uses a cutoff of 0) in a community with 39 good agents performs even better than a single agent using a cutoff of 1 (having an MII of approximately 2.411 as opposed to 2.405 in simulations, although the difference is not statistically significant). However, one can also see from Figure 4 that the presence of a higher percentage of good agents leads to a better outcome for every type of agent. Both good and bad agents perform better when there are more good agents present. If every single agent is bad and defaults a large fraction of the time, this is globally sub-optimal. These results are an example of what is known as the "free-rider" problem in game theory.

4.5 Discussion

The results presented here indicate that cutoff functions represent a useful space of policies to study in the SPIRE framework. The smooth curve seen in Figure 3 indicates that there exists a cutoff that globally maximizes mean individual income in homogeneous agent communities. It is also likely that there is a single best cutoff for an agent to use in any given community. While this cutoff may not represent the best policy if a different model of the evolution of reputations is used, the linearity of the reputation functions ensures that the cost of defaulting to an agent is independent of the value of the agent's individual reputation. Therefore, the decision of whether to default or not should be independent of all aspects except the difference in outside-offer and group-related task values.

Experiments such as the ones described here can be used to implement policies to address intention reconciliation. From Figure 3, we can surmise that the cost to the group of defaulting is about 2. As policy-makers, we can institute a "default fee" of 2 to be paid by an agent when it defaults, and do away with individual reputation. The effect of this policy should be to turn all agents into approximately $\mathbf{CF}(2)$ agents. The advantages of this policy are that it achieves high revenue, is stable, and avoids the free-rider problem. The disadvantage, of course, is that it limits the ability of creative agents to figure out better solutions, particularly if the environment changes.

5. LEARNING IN THE SPIRE MODEL

The performance of an agent using a fixed policy is dependent on two factors — the environment, represented by parameters such as the weights of **IR** and **GR**, and the agent community, represented by the size of the group and the policies being used by other agents. Experiments show that there is no single policy an agent can follow that optimizes its performance across a range of different communities and environments in the SPIRE model. Further, designers of multi-agent systems in the real-world cannot run simulations for all the different kinds of environments in which their agents are expected to operate. It is necessary for agents to adapt to their environments and communities to optimize the utilities they receive. This section investigates adaptive agents that learn policies based on the environments and agent communities they are part of. The algorithm we present can be used successfully by a single agent to learn near-optimal cutoffs when the other agents are playing fixed strategies. It also scales reasonably well to situations in which multiple agents are using the same algorithm to learn how to behave.

5.1 Learning the Cost of Defaulting

Agents learn cutoffs by keeping running estimates of the values that will be received in the future from defaulting and not defaulting. To do this, an agent keeps track of its behavior over the last n time steps. An agent maintains estimates of e_{def} and e_{nodef}, its expected total rewards in the n time steps after defaulting and not defaulting, respectively. At each time step, it looks at the action taken n steps earlier, computes the reward earned over the last n steps, and updates the relevant estimate. The estimate is updated incrementally using some learning rate η.

The estimates are used for decision-making as follows: at a given time step, an agent defaults if the difference between the value of the outside offer (O) and the value the agent receives from the group-related task (T) is greater than the agent's estimate of the difference in utilities it will receive over the next n steps if it defaults or if it does not default. There is an exploration probability ϵ associated with each decision. With probability $1-\epsilon$, if $e_{def} - e_{nodef} < O-T$ then the agent will default, otherwise it will not default. With probability ϵ the agent will take the contrary action, defaulting if $e_{def} - e_{nodef} \geq O - T$ and not defaulting otherwise. For the experiments reported here, we use $\epsilon = \frac{1}{t^{0.3}}$ where t is the number of steps since the beginning[4]. ϵ decays over time, which means that the policies agents use become less exploratory, with agents choosing the action their estimates suggest is optimal more as time goes on. This process is especially important because the agents are learning their estimates over n steps in the future, and if intermediate actions are not optimal the estimates are less accurate.

The selection of n, the number of steps that an agent should look-ahead in the future and η, the learning rate to use, raises some interesting issues. Theoretically, the cost of defaulting at a given time step should extend indefinitely into the future. Therefore, by restricting n to some finite number, we introduce some bias into the estimate, which will result in the estimate being lower than the true cost. However, as we increase n, although the mean should become closer to the true cost, the variance of the estimate goes up, making the estimates rapidly less useful as we increase n beyond a point. Using a smaller value of the learning rate η could reduce the variance of the estimates, but that comes at the cost of slower learning, which is not desirable, especially in real-world situations. After extensive experimentation, we decided to use parameter settings of $n = 5$ and $\eta = 0.02$ as default values in the experiments presented here.

5.2 Single Agent Learning: Empirical Results

We performed experiments in which a single agent was learning in a community of agents using fixed cutoff values. Three of these experiments were in communities of 10 agents, 9 of whom were using the same cutoff, and the other three were with communities of two agents, one of whom was learning while the other used a cutoff of 2. The results are summarized in Table 1. The categories for best MII and cutoff (used to achieve the best MII) are based on replicating the experiments with 10 or 2 agents but using one agent with a different cutoff in the set $\{0, 1, 2, 3, 4\}$ instead of the learning agent. This method should give a good approximation to the best MII that can be achieved by a single agent in the given community.

These results are representative of a series of such experiments, in all of which the learning agent achieved performance similar to the best performance achieved by a cutoff agent using an integral cutoff. The changes in default percentage indicate that the agent is learning and adapting its behavior to the community it finds itself in. For example, when 10 agents are in the community and the other 9 are using cutoffs of 0, it is advantageous to use a very low cutoff; there is little to be gained from not defaulting very much, because an individual's effect on group reputation is small. However, in a two agent community in which the other agent is using a cutoff of 0, it may be better to have a slightly higher cutoff, because an individual's behavior can have a greater impact on group reputation. Another interesting aspect of the results that also holds for the multi-agent learning experiments is that agent behavior is not significantly affected by the choice of the initial estimate of the cost of defaulting (for reasonable estimates). We used an initial estimate of 0 for all experiments reported here.

5.3 Multi-Agent Learning

An algorithm that enables a single agent to learn near-optimal behavior in communities of agents that use fixed policies is important, but agent designers would also like to enable the algorithm to be successful when other agents could also be learning. It is important to design and evaluate learning algorithms with this goal in mind. An important component of the cutoff-learning algorithm we present in this paper is the learning rate, η. If this learning rate is kept constant, the algorithm should gradually forget past history, giving more recent occurrences more weight in its estimate of the cost of defaulting. On the other hand, if η decreases over time, for example with a function like $\eta = 1/t$, then each occurrence in the past will have equal weight at any time step. The former method should be more suited to non-stationary problems like the multi-agent learning problem.

To examine the effects of learning rates that stay constant versus those that decrease over time, we ran experiments in which all 10 agents used the same learning algorithm. The results of these experiments were surprising. The performance of communities of 10 agents that are all learning was the same for both time-decreasing and constant learning rates. In both cases the agents accrued a mean individual income of 1.98 ± 0.02 over the last 2000 time-periods of a 10000 time-period simulation, although the agents using a constant learning rate defaulted significantly less (on approximately 87% as opposed to 98% of opportunities). While the different default rates show that the agents are learning different types of behavior, the performance of the agents using a constant learning rate is surprisingly poor. This poor performance can probably be attributed to the variance of agent estimates of the cost of defaulting at any given time. Because there are 10 agents in the framework, the estimates that each agent has at any given time will probably be different from the estimates of a number of other agents by a fairly significant amount, and thus it is possible that the agents do not converge to a more optimal default rate.

In support of this hypothesis, the results in communities of just 2 learning agents (in which there will be less variance between agent estimates) are significantly better, with agents achieving mean individual incomes of 2.14 ± 0.02 and defaulting on approximately 69% of opportunities.

We also experimented with scenarios in which 5 agents took part in a 12000 time-period simulation, and each used a fixed cutoff (of 2) for a certain number of time steps and then started learning (the settings simulate environments in which agents change their behavior over time, or agents are replaced by different types). The results of this experiment are reported in Table 2, where agent 1 learns from the beginning, agent 2 uses the fixed cutoff for 2000 time periods and then learns, agent 3 uses the fixed cutoff for 4000 time periods and then learns, and so on. We used 12000 as the number of time periods to give agent 5 some time to learn. Agents used a fixed learning rate of 0.02.

Each income is reported to within a confidence interval of ± 0.03. The agents all behave similarly, defaulting on between 77.3% and 79.1% of opportunities, in spite of the fact that they started learning at different times. The fixed

[4] Initial experiments suggested this was an appropriate function to use.

Table 1: Single Agent Learning

# Agents	Cutoff	Best MII	Best MII Cutoff	%defs	MII of Learner	% defs of Learner
10	0	1.99 ± 0.02	0	100	1.97 ± 0.02	84.2
10	2	2.26 ± 0.02	1	65.3	2.25 ± 0.02	81.4
10	3	2.34 ± 0.02	0	100	2.34 ± 0.02	83.9
2	0	2.06 ± 0.02	1	66.9	2.05 ± 0.02	71.1
2	2	2.22 ± 0.02	1	66.4	2.21 ± 0.02	69.0
2	3	2.26 ± 0.02	1	66.8	2.25 ± 0.02	66.9

Table 2: MIIs for Stepped Learning

Ag 1	Ag 2	Ag 3	Ag 4	Ag 5
2.05	2.08	2.06	2.08	2.06

learning rate allows them to forget the past and adjust to the present state of the environment, which is crucial for learning in nonstationary environments. Agents using a time-decreasing learning rate default more in similar experiments, because it is advantageous for a learning agent to default more earlier in time, when other agents using a higher cutoff can be exploited.

6. CONCLUSIONS AND FUTURE WORK

This paper makes four main contributions:

- a new, simpler framework for studying intention reconciliation that allows for mathematical specification and provides a basis for the study of learning;
- an analysis showing that a simple "reactive" cutoff policy can be expected to perform as well as more complex policies that look ahead;
- an empirical investigation showing that an intermediate level of "social consciousness" leads to optimal results and demonstrating the free-rider problem;
- an algorithm for learning the optimal cutoff value based solely on observed task values.

There are a number of ways in which to extend the study in this paper. The overall framework and the evolution of reputations is designed to be robust over a wide range of environmental settings, but further empirical investigation would be useful.

One way to extend our model is to allow replacements, where an idle agent can substitute for an agent that defaults on a group task. Another is to allow agents to enter and leave the group over time. It would be interesting to see whether or not these modifications significantly change the findings presented here.

The most significant bottleneck is the difficulty of multi-agent learning. Progress is being made in that area, with the most promising approach being to learn mixed strategies [1]. From a strategic point of view, it is unlikely that mixed strategies would be required in our domain. From a learning perspective, however, mixed strategies may be more robust to changes in the environment. Our model provides an interesting testbed for new techniques in this area.

7. ACKNOWLEDGEMENTS

This research has been supported by NSF grants IIS-9978343 and IRI-9618848. This paper reports on work that formed part of the first author's undergraduate honors thesis at Harvard University.

8. REFERENCES

[1] M. Bowling and M. Veloso. An analysis of stochastic game theory for multiagent reinforcement learning. Technical Report CMU-CS-00-165, CMU, 2000.

[2] M. Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.

[3] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of AAAI*, 1998.

[4] A. Glass. Creating socially conscious agents: Decision-making in the context of group commitments, 1999. Senior Honors Thesis, Harvard College, Cambridge, MA.

[5] A. Glass and B. J. Grosz. Socially conscious decision making. In *Proceedings of AGENTS*, 2000.

[6] B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86, 1996.

[7] B. J. Grosz, S. Kraus, D. Sullivan, and S. Das. The influence of social norms and social consciousness on intention reconciliation. *Artificial Intelligence (ICMAS-2000 Special Issue)*, 2002. To appear.

[8] M. L. Puterman. *Markov Decision Processes.* John Wiley and Sons, 1994.

[9] T. Sandholm and V. Lesser. Advantages of a leveled commitment contracting protocol. In *Proceedings of AAAI*, 1996.

[10] S. Sen and A. Biswas. Effects of misconception on reciprocative agents. In *Proceedings of AGENTS*, 1998.

[11] L. Shapley. Stochastic games. In *Proceedings of the National Academy of Sciences of the USA*, 1953.

[12] D. Sullivan, A. Glass, B. Grosz, and S. Kraus. Intention reconciliation in the context of teamwork: an initial empirical investigation. In M. Klusch, O. Shehory, and G. Weiss, editors, *Cooperative Information Agents III.* Springer-Verlag, Berlin, 1999.

[13] D. Sullivan, B. Grosz, and S. Kraus. Intention reconciliation by collaborative agents. In *Proceedings of ICMAS*, 2000.

[14] V. Teague and L. Sonenberg. Investigating commitment flexibility in multi-agent contracts. In *Proceedings of the 2nd GTDT Workshop*, 2000.

[15] O. Vrieze. Stochastic games with finite state and action spaces, 1987. No. 33. CWI Tracts.