

Communication, Credibility and Negotiation Using a Cognitive Hierarchy Model

Michael Wunder
Rutgers University
mwunder@cs.rutgers.edu

Michael Littman
Rutgers University
mlittman@cs.rutgers.edu

Matthew Stone
Rutgers University
matthew.stone@rutgers.edu

ABSTRACT

The cognitive hierarchy model is an approach to decision making in multi-agent interactions motivated by laboratory studies of people. It bases decisions on empirical assumptions about agents' likely play and agents' limited abilities to second-guess their opponents. It is attractive as a model of human reasoning in economic settings, and has proved successful in designing agents that perform effectively in interactions not only with similar strategies but also with sophisticated agents, with simpler computer programs, and with people. In this paper, we explore the qualitative structure of iterating best response solutions in two repeated games, one without messages and the other including communication in the form of non-binding promises and threats. Once the model anticipates interacting with sufficiently sophisticated agents with a sufficiently high probability, reasoning leads to policies that disclose intentions truthfully, and expect credibility from the agents they interact with, even as those policies act aggressively to discover and exploit other agents' weaknesses and idiosyncrasies. Non-binding communication improves overall agent performance in our experiments.

Categories and Subject Descriptors

F.1.1 [Computation by Abstract Devices]: Finite Automata; G.3 [Probability and Statistics]: Markov Process; I.2.0 [Artificial Intelligence]: Cognitive Simulation; I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—*Intelligent Agents, Multiagent Systems*

General Terms

POMDPs, Cognitive Hierarchy, Repeated Games

Keywords

Agent Communication, Game Theory, Multiagent Learning

1. INTRODUCTION

When agents repeatedly face the same simple strategic problems, we can expect them to exhibit *equilibrium* behavior—where each acts to optimize his outcome given how the others will act; see Leyton-Brown and Shoham [8] for a compu-

tational introduction. For rational agents, such settings offer the information and computation needed to deliberately work out a strategy that takes others' choices into account. But even agents that act by trial and error will reach equilibria in such settings, through the reinforcement provided by patterns of past play [10]. Indeed, in repeated simple strategic interactions, we do find equilibrium behavior in people and many other organisms.

It is different for situations which do not obviously call for strategic reasoning, for situations which are novel or complex. Here, it becomes unlikely that agents will enjoy the information, computation or history that make equilibrium play feasible. Instead, we expect agents to exhibit heuristic, approximate or biased solutions to the problems of interaction [1, 9]. Think of a lovers' quarrel, a game of chess, or a corporate acquisition: these are interactions people muddle through, not interactions they solve. In interesting domains, agents will have to go beyond first-principles reasoning about interaction to model the heuristic basis of their partners' decisions: the gut reactions, the resource limits, and the background culture of practices and expectations that actually lead to choices. Agents may also have to work explicitly to create interactions where their own and their partners' limited reasoning is effective. Such thinking is the mainstay of human social life: not only in the trust, rapport and relationship we can achieve in alignment with one another, but also in the circumspection, prudence and vigilance we otherwise maintain.

This paper presents a set of simulation results designed to explore the diverse reasoning that might be required for interacting agents in realistic situations. Our basic tack is to approach interaction not as strategy but as a more constrained planning problem for sequential action under uncertainty. In particular, we use the computational framework of partially-observable Markov decision processes (POMDPs [6]) as a common substrate for our agents' reasoning. Our POMDP framework captures the key features of interactive reasoning by simulating the fixed decision making of relevant possible opponents. We can represent *background expectations* and *heuristic choices* by simulating opponents that meet specific expectations, or follow specific heuristics. We can characterize *limited strategic reasoning* through a cognitive hierarchy model—simulating more sophisticated opponents that tune their behavior to simpler opponents [3]. And finally, we can investigate *adaptive response*, because the POMDP framework describes the tradeoff between acting to achieve immediate payoffs and acting to reveal qualities of opposing players that can be exploited in further

decision making.

The method we introduce applies POMDP solving techniques to find the best response to a given population of strategies. A new population is built by combining the old one with the new computed strategy. As this process is repeated and the population grows according to a predefined Poisson distribution, earlier strategies recede in importance somewhat. In this way, our proposed algorithm repeatedly computes new strategies and attains more experience against more sophisticated agents, as it converges toward some policy that will yield the same policy in response. This procedure therefore determines a *cognitive hierarchy model*, as explained more fully in the Section 2.1.

We work specifically with the classic prisoner’s dilemma game, iterated for a small number of rounds with an opponent with limited, heuristic, or biased decision making. In this setting, our agent will do best by defecting against uncooperative or credulous opponents, and by engaging with reciprocating opponents who respond to cooperation with cooperation but defection with defection. Thus, our agent must learn how its opponent plays, and signal how it plays, as it accumulates payoffs in initial rounds. We sometimes give agents the opportunity to explain themselves: to offer non-binding promises and threats describing the expected course of future interactions. Our simulation results show that agents in this general setting can achieve collaboration in some cases, but that they face a difficult problem in establishing trust with a reciprocating opponent while guarding against exploitation by uncooperative opponents. Communication can help, by allowing agents to establish trust by truthfully describing harmful actions they must take to avoid leaving themselves vulnerable to exploitation. Specifically, once the model anticipates interacting with sufficiently sophisticated agents with a sufficiently high probability, reasoning leads to policies that disclose intentions truthfully, and expect credibility from the agents they interact with, even as those policies act aggressively to discover and exploit other agents’ weaknesses and idiosyncrasies. The results offer a suggestive link between the challenges agents face in connecting with one another and the constructs that people—or agents—can use to pursue those connections.

2. RESEARCH SETTING

Traditional learning methods find it difficult to learn in games where other agents are also learning. The algorithms often require a static environment and the presence of other learning agents invalidates this assumption. One commonly cited example is the rock–paper–scissors game, where each one of the three moves dominates one other. In this case, crudely adapting agents will circle each other indefinitely, as each reacts to a model of the other that is out-of-date.

One lesson is that acting effectively amid other learners requires a new way to generalize from past experience. It is not enough to assume that other agents will act as they have in the past. Others may think strategically, and arrive at behavior that differs from anything that they have done so far. To prepare for this, agents must second-guess one another, and attempt to model their opponents’ possibly creative decision making. Such reasoning strategies can lead to good performance against learning agents, even in problematic cases like rock–paper–scissors [5].

In general, we could approach this second-guessing through the symmetric strategizing of game theory [8],

through open-ended general models of agents’ reasoning [5], or through a range of further simplifying assumptions. The cognitive hierarchy model offers one such simplification; it offers a framework for bounding strategic reasoning by assuming that agents have heuristic limits on the degree to which they are willing to second-guess one another [3]. The model is intuitively appealing, because it captures the empirical observation that players are boundedly rational, with limited amounts of working memory, and furthermore believe that others also have bounds on their rationality. Moreover, as we shall see, it has a straightforward implementation using off-the-shelf planning techniques.

2.1 The Cognitive Hierarchy Model in Games

A cognitive hierarchy model is defined in terms of a set of base policies and a series of levels of sophistication describing agents’ possible strategic reasoning. The base policies involve no strategic reasoning; they simply prescribe a distribution over actions to perform in each state. The least sophisticated agents in a cognitive hierarchy, the 0-step agents, are those that directly follow one of the base policies. We assume an initial distribution over the 0-step agents.

More sophisticated agents at level $k + 1$ in the hierarchy perform best-response reasoning, on the assumption that other agents are drawn from a distribution of agents at level k or lower. Thus, agents at level $k + 1$ have no model of players doing more than k steps of thinking. For example, a 1-step agent second-guesses the 0-step agents by planning a best-response to their distribution of strategies.

In general, k -step players must plan their responses based on a distribution over the sophistication of their opponents. A simple assumption is that their information about less sophisticated agents is correct; their “mistake” is to ignore the possibility of more sophisticated agents. Formally, if the actual proportion of h -step players is $f(h)$, a k -step player will believe the proportion of h -step players to be $g_k(h)$, where $g_k(h) = 0 \forall h \geq k$ and $g_k(h) \propto f(h)$ otherwise. Finally, as k increases, we assume that k -step players become exponentially less and less likely, so this estimate converges to a fixed distribution. Essentially, the difference between players doing k and $k + 1$ steps of thinking will shrink as k increases and the proportion of agents reasoning more than k steps get smaller. Under these conditions, the players should converge towards some strategy. However, we might not expect this algorithm to end up at a Nash equilibrium. The agent must respond to the original naive strategies as well as the more sophisticated ones, as a result of keeping some of them in its model of the environment.

The distribution $f(k)$ is the main component of the cognitive hierarchy model, and is determined by the parameter τ corresponding to the average k in the population. Camerer et al. [3] argue that $f(k)/f(k - 1)$ is proportional to $1/k$, resulting in $f(k)$ as the Poisson distribution, so that $f(k) = e^{-\tau} \tau^k / k!$. This distribution has a number of advantages: it is simple to compute as it has only one free parameter, τ , and it closely fits empirical observations.

One example where this model is useful is the Keynesian beauty contest [7] where players score highest by picking the contestant that others value most. In the p -beauty contest, numbers are substituted for beauty contestants. Each of N players chooses a number $0 \leq x_i \leq 100$. The average of these numbers scaled by p specifies the winning number. The player who chose closest to the winning number wins a

fixed prize. The Nash equilibrium for this game is for everyone to choose 0 when $p < 1$, but people do not reach this equilibrium immediately. A cognitive hierarchy model can explain these results in terms of players' expectations over how deep the average reasoning will be, with increasingly sophisticated agents reaching increasingly low estimates of others' likely bids.

2.2 POMDPs

Partially observable Markov decision processes, or POMDPs, are a mathematical model for decision making under uncertainty and have been used productively to define and solve for optimal policies for agents [6]. A POMDP model consists of sets of states, actions, observations, and probabilistic functions relating them.

The problem of computing a policy that plays optimally against a finite-state opponent chosen from a finite set can be cast as a POMDP, consisting of an underlying MDP $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$ and set of observations \mathcal{O} . Here, the states \mathcal{S} of the POMDP are the states within the possible strategies. The actions \mathcal{A} are the choices of the agent and transition function T is determined by the opponent strategy. Rewards R are a property of the game and depend on both players' actions. The observations \mathcal{O} are the opponent's observed responses. When there is communication in the game, the receipt of messages between rounds map to additional observations in separate states. An agent starts out in state s against strategy i , but this state is unknown to the agent.

Through repeated interaction over the length of a POMDP, an agent computes its belief state b , which is a cross product of all possible strategies and the current state in each such strategy. As the agent traverses the graph defined by the opponent's strategy, it achieves greater certainty about the identity of that opponent as well as the proper response. In other words, each new observation o brings b more in line with the correct state within the actual strategy of the opponent. Solving methods balance the value of this information with well-rewarded actions based on b .

A POMDP solver available on the web (www.pomdp.org) presents a straightforward protocol for specifying a POMDP, including states, transitions, actions, rewards, and observations [4]. In combination, these five inputs constitute a POMDP. The solver uses several state-of-the-art exact algorithms to calculate the optimal policy and values for each state depending on the actual probability for these states. If the solver can construct an optimal policy within reasonable time bounds, it outputs this strategy the form of a finite automaton, which is easily incorporated as another entry in the cognitive hierarchy for future optimizations.

To capture a cognitive hierarchy with a POMDP solver, we run the solver on a model that describes interaction with a distribution of 0-step agents. The solution describes a 1-step agent. From here on out, whenever we derive a solution agent for step k , we can add the agent to the distribution of agents with weight given by $h(k)$, and repeat the solution process, yielding the automaton for step $k + 1$. The procedure can be iterated indefinitely.

3. TRADEOFFS FOR REASONING

Games with cooperative outcomes and an incentive to cheat, like the prisoner's dilemma, are good settings for exploring ways to encourage and enforce cooperation between agents with conflicting interests. We consider solutions to

finite iterated prisoner's dilemma problems under a cognitive hierarchy model. We begin with a range of base policies, some of which reward cooperation and punish defection, and some of which do not.

The best response against these base agents amounts to a tradeoff. Defecting offers an opportunity to discover what kind of opponent one faces, and perhaps to exploit them. But it forgoes opportunities to cooperate with reciprocating agents, against which cooperation is the best outcome.

More sophisticated agents eventually face different tradeoffs. They must not only be able to size up base agents, but they must be able to recognize and interact appropriately with agents that carry out limited strategic reasoning. Depending on the prevalence of the different levels of reasoning in the environment, a more sophisticated agent may be forced to surreptitiously masquerade as a reciprocating base agent in order to elicit good behavior from less sophisticated agents.

3.1 Communication and credibility

Consider what happens if agents can exchange non-binding threats and promises that do not impact the payoffs of the game. In equilibrium analyses, 'cheap talk' of this form has an effect on the outcome of a game only under certain constraints. In particular, the type of the sender agent must make a difference to the receiver, and the sender and receiver cannot diverge in preferences. While for us the first condition is clearly met, the second is not because the agents have incentive to defect from the cooperative outcome, and therefore to *deceive*.

In the cognitive hierarchy model, however, the effect of cheap talk depends not only on the game itself, but also on the behavior of the base agents and the tradeoffs made by agents of different levels of sophistication. We include reciprocating agents that describe their future intentions honestly. Meaningful communication cannot occur without some agent in the population who engages in enough truth telling to make it worthwhile to believe the agent.

Best responders to these base strategies then have the option of demanding integrity rather than demanding reciprocation, or in addition to it. Such agents provide benefits to truth-tellers, and punish deceivers.

More sophisticated agents may not only call for integrity but realize that they must behave with honesty if they are to cooperate with the less sophisticated agents that demand integrity—the analogue of masquerading as the kind of base agent that these agents cooperate with. In the end, such agents communicate successfully, truthfully sharing the information that they are hard-nosed players that nevertheless intend to reciprocate with one another. This is a way to get meaningful communication despite the nonaligned interests of the communicators.

3.2 Learning and teaching

Reasoning in the cognitive hierarchy contains elements of strategic teaching [2], as a side-effect of the iterated planning involved in solving POMDPs. In essence, by computing the optimal series of actions given a certain 0-step distribution over opponents, the output 1-step solution will act to help train the 2-step generation's computed policy. The next policy shaped in part by the interactions of the previous one will be guided towards the 0-step reciprocal strategies. This effect will feed on itself, and the 2-step agents will in turn

guide the 3-step generation even more strongly. Another property of the algorithm based on the cognitive hierarchy viewpoint is that the proportion of sophisticated agents that each new k -step POMDP contains gradually rises for all values of τ , leading to a situation where the majority of agents contain some level of sophistication against several classes of opponents.

Other research has looked into adaptive methods to finding best actions, notably experience-weighted attraction (EWA), which itself blends two other techniques, relying on reinforcement and belief updating [2]. Our iterated method attempts to compute an exact policy for the entire game period and finds the optimal series of moves for the total payoff. Another difference is that EWA tries to adapt to changing players as the game progresses. This algorithm yields stationary strategies and the contribution of each strategy to an agent’s model changes as k and τ increase. Our model considers other strategies to be stationary multi-step decision rules and must adapt to changes in the relative proportions of such rules.

4. EXPERIMENTS AND RESULTS

The following experiments were run with Cassandra’s POMDP solver from www.pomdp.org [4]. The game is a six-turn version of prisoner’s dilemma, with the following payoff matrix for player 1. Player 1’s actions are on rows, player 2 on the columns. C stands for the cooperate action and D for defect:

$$\begin{bmatrix} P1 & C & D \\ C & 3 & 0 \\ D & 4 & 1 \end{bmatrix}$$

In this game, a player is always better off choosing D, which leads (D,D) to be known as the Nash equilibrium for this game. However, these values tend to create cooperative behavior when playing is repeated. There is also a discount factor of 0.95 for each successive state. Where communication is present, the message-exchange step is a separate state, but for purposes of comparing relative scores the discount will be excluded.

One useful feature about this solver that is convenient for these experiments involves the output values for each belief state. This data is output along with the policy graph finite automaton solving the POMDP. While the solver provides one policy based on the starting state specified in the program, policies for different starting belief states can be calculated without running the solver again. This way, the policies can be quickly recomputed without creating new files for each desired combination of belief states.

4.1 Iterated Prisoner’s Dilemma Without Communication

Imagine that you are told to play IPD for six turns against a single player selected at random from a population. The majority of this population will be unknown. Some, but not all, of your possible opponents will consist of the following ratio of four strategies, defined as the 0-level. Three of the four 0-level strategies are non-reciprocating, including a strategy that always cooperates, one that always defects, and one that simply picks an action randomly, with each action equally likely. A fourth strategy is reciprocating, meaning that it pays to cooperate with it as cooperative actions

will be rewarded with cooperation, while defections will tend to be punished with defections. A player has equal chance to meet each of these strategies. We might consider them to represent simplified personalities that exist in the world. There exist the altruist, the mean individual, the crazy, random strategy and the reasonable, reciprocating player. The ratio of three to one is derived from the relative costs of defecting against the two classes of agents. On average, the payoff of defecting against a non-responsive player is +1, while the average penalty for defection against a fully reciprocating agent is no smaller than -1.5 in the long run (-3 every other turn). A minor alteration is that a small percentage around 20% of the reciprocating population starts off defecting while the rest cooperate. While this change does not have much effect on the actual policies, it makes exploration easier.

Examples of reciprocating agents include Tit-for-Tat (TFT) and Pavlov. TFT starts by cooperating and then simply returns the action of the other player from the previous round. After the first round, Pavlov will cooperate if the players play the same action in the previous round, and otherwise it defects. For the reciprocating player we have chosen to use the Pavlov strategy, for several reasons. First, Pavlov achieves a subgame equilibrium in self-play in repeated settings. It has a credible deterrent and is somewhat more robust than TFT where some degree of noise is present, as two TFT agents can easily start alternating defections. It is also an evolutionarily stable strategy. Another reason is as there are only a short number of rounds of play available, it is necessary for a strategy to achieve coordination quickly. TFT, while ideal as a deterrent, adds uncertainty when players find themselves in mutual defection. A player will have to cooperate for two more turns before finding out whether the opposing strategy matches TFT, while against Pavlov it will be known the following turn.

It is uncertain what a player should do in a situation like this one, where players other than these given strategies will be part of the population alongside the known participants. These other players should be considered to have some reasoning capacity, and are trying to figure out what to do also. Someone might start by saying that it would be nice to cooperate with the reciprocating player, assuming that the same actions are observed. However, the observed actions are the same for the always C agent, so it would also be nice to know which one we are facing, as there is substantial reward for having this knowledge. A player can only find out this information by testing the other player with a defection move. The problem comes when the player starts thinking past this point to how more advanced players will view the game. If a defection is sent out the first round, it will be indistinguishable from the constantly defecting agent. So, perhaps it would be better to wait a certain number of turns, but during this waiting period a player is giving up valuable opportunities. It would appear that this type of game is suited for cognitive hierarchy analysis, where the four given types of player constitute the 0-step thinker in that model.

In order to implement our version of the cognitive hierarchy model, we include each k -step agent according to the predefined Poisson distribution. The primary parameter for the distribution $f(k)$ is τ , and it will be varied across trials, also determining the actual number of 0-step players. Once a policy against the initial distribution is computed, it is called the 1-step policy. It is included in the set of strate-

gies in the proportion given by $f(1)$ for the second iteration of POMDP, which will yield the 2-step agent. However, the estimated population does not yet include agents with step greater than 1, so at every stage only policies computed with step size $< k$ will be present. The distribution of strategies in round k will be attained by renormalizing the contribution of the original strategies given by $f(0)/\sum_{i=0}^{k-1} f(i)$. The process is then repeated to find the 3-step policy, and so on.

Some simple calculation will show that even with low levels of reciprocating agents, it is worthwhile to engage in cooperation, provided that a player can become fairly confident about the opponent after several testing rounds. Figure 1 shows the number of testing moves required to distinguish a reciprocator from a completely random agent, over varying prior ratios for the reciprocating agent. If the agent fails the test, the other agent is clearly random and exploitable.

When Pavlov makes up 25% of the initial arrangement, the first computed policy begins by testing the field with a single defection, and then follows the rule set by Pavlov. Namely, if the other agent has defected also, cooperate next. Otherwise, defect again. This policy can be seen in Figure 2. This combination accomplishes two things: it forces the reciprocating agent to respond with a punishing defection on the following round, and it reveals some of the non-responsive agents such as the always cooperating agent. Since there is only one correct sequence of actions for this test, the random agent gives itself away with increasing likelihood. If the opposing agent passes the test, the 1-step player is then safe to pursue mutual cooperation at least until the final round, when there is no consequence for taking the more profitable defection move.

This strategy is placed into a new POMDP at a proportion derived from the current τ . The 2-step policy is then computed. See Figures 2-5 for the first four steps of reasoning computed in this manner, with $\tau \leq 2$. To read the finite automata graphs, start at the state with the double circle. The action inside the circle, whether C or D, describes the agent's action. The arrows exiting the circle are the actions played by the player facing this agent. The agent behavior can then be tracked until the game ends.

Another feature of the policies with small k is that unwarranted defections, such as those that occur without a prompting test defection, are classified as non-responsive. This feature has consequences for the next round of policy computation, because the condition of one or more initial cooperation moves must be met to warrant future cooperation from the population of 1-step policies. As Figure 3 shows, the 2-step policy finds it beneficial to wait a turn before testing the opposing agent, as well as to watch for the pattern defined by the reciprocating 1-step agent. For lower values of τ this process continues until k is about 4, where the number of previous $(k - 1)$ -step agents are too small to consider. The final policy of this particular sequence for $\tau \leq 2$ in Figure 5 shows how waiting three turns is the best policy, followed by constant defection if it has not been tested by that point.

For intermediate $\tau > 2$, the process gives a strategy where agents find that it is only worth cooperating for a single round if the other agent is very forgiving, due to the wide range of strategies present in the population. In the highest values of τ measured, about 3.5 and 4, a new policy arises in response to the dominant one, which acts very obediently in order to receive its one cooperation outcome from the

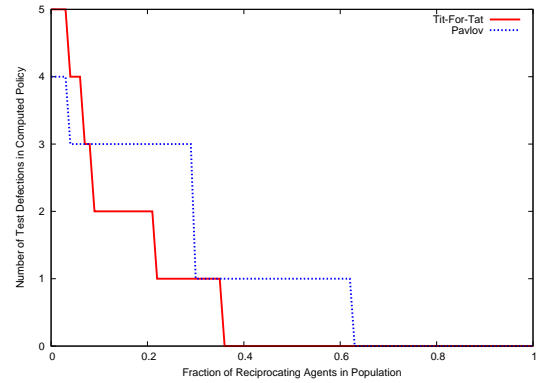


Figure 1: Number of testing rounds in a game between random agent and reciprocator required to reduce probability that the opponent is random. Plotted over the true probability of reciprocation.

previous strategy.

Figure 6 demonstrates the relationship between τ and the overall population's score. We see a slight rise for low values of τ , with a peak around 2 followed by a steep dropoff. In general players do better within a population by reasoning more, but the population suffers when too much reasoning is done because the cooperation unravels from the final turns.

4.2 Iterated Prisoner's Dilemma With Communication

The same method applied to games with communication proceeds somewhat differently. Here, the players are able to send a message before odd-numbered turns, which consists of a two-bit string. This message is meant to correspond to the next turn consequence to the opponent for its choice of action in the current round. That is, the first string describes the sender's action resulting from a receiver's cooperate action, and the second bit of the string describes the response to the receiver's defection action. Thus, two rounds are required between messages. This conditional structure can also be applied to games with more than two actions.

These messages will gain their meaning from the agents that intend to enforce the truth value of messages they send, even if only consisting of a minority of the total population. An important point concerning these messages is that only two of them actually contain evidence that the sender will reciprocate. A message containing the same action for both conditions says either that the receiver's action will be ignored in the decision for the next turn, or that the sender does not mean what it is saying. In either case, these messages express that the sender is an unreliable communicator and does not deserve reciprocation. This condition holds as long as there are no reciprocating agents that fail to communicate correctly.

The message stage opens up more possibilities for the initial 0-step agents to play. There are four new actions between every second round that are seemingly unrelated to the underlying strategy. The communication strategies mirror their corresponding actions. That is, the virtuous reciprocator agent follows a Pavlov strategy that communi-

cates that strategy truthfully. The random agent sends random messages. The all-cooperator and all-defector attempt to blend in with the truthful population by restricting the messages to the Tit-for-Tat message, which states that an action will be met with the same action on the following turn. Looking more closely, the naive cooperator demonstrates even greater naivete in the hope that an empty threat will induce cooperation. Since this message is the one most commonly used by the Pavlov strategy, it also makes sense that an all-out defector would send it, as this lie has the biggest chance of success. In general, the games with and without communication were made as close as possible in the sense that the same starting population was used in both. While communication is itself a big difference, it is not immediately obvious that it will have much impact on the actual results.

The population of k -step agents follows the same distribution as the non-communication case. When τ is varied, the fraction of each policy changes. As the policy computation is repeated, initial differences in the distribution over k will lead to different policies, and in substantial divergence in later rounds.

In the communicating population, the 1-step policy computed on this 0-step distribution defects right away on the non-information messages, as only the random agent sends such messages. It also becomes very sensitive to the truth content of the messages. Senders that do not conform to their own messages by following through on their promised consequences are weeded out by the second turn. To complete the evaluation, the agent starts off by testing the results of defecting, and if the action is unpunished continues that exploitation. Otherwise, reciprocation continues pending further adherence to the truth contained in the messages. It is important to note, however, that this policy does not meet its own standards. Since there is no opposing agent that will exact a consequence for reneging on promises, it does not pay attention to the messages it sends.

The 2-step agent does send meaningful messages, and sticks to them until an opposing promise is broken. The presence of the 1-step agent is enough to select for this property. Furthermore, because the computed policies find truth enforcement to be a better indicator for future reciprocation, there is no need to seek the cooperative action from the beginning. Therefore, the 2-step agent is free to test by defecting right away, and checks that the other agent will follow direction by sending a specific message that corresponds to the message a Pavlov agent would send in that scenario. Specifically, the message states that "I will defect next if you cooperate this turn, and vice versa". Any logical agent could not resist this temptation to defect, and so if the other agent matches this behavior, mutual cooperation can begin on the following turn. If the other agent starts off with the Tit-for-Tat message, the correct sequence will be cooperate/defect for the opposer, and defect/defect for the computed 2-step policy. Mutual cooperation can then ensue in the third turn, as long as the messages now contain the Tit-for-Tat instruction.

An agent dealing with the 2-step agent will be led to truthful communication by the cascade of reasoning suggested in Section 3.1. The 2-step agent finds the cooperative outcome with the Pavlov agent as well as other sophisticated strategies. Second, it tests the other agent's messages for accuracy. Third, it enforces this credibility test with a penalty of

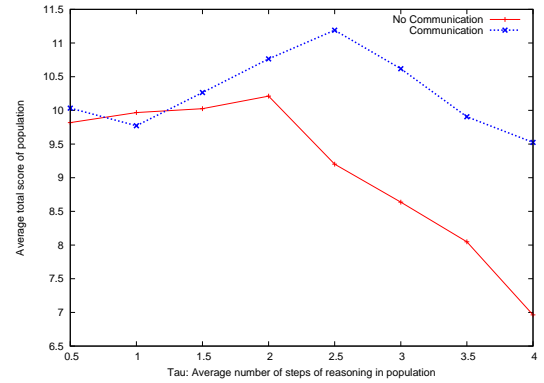


Figure 6: The total score of a population of agents, weighted by proportion of population, as the average level of reasoning changes

endless defection for failure. Finally, it tells the truth itself, at least until the last round when future actions cease to matter.

Across the board, as τ varies, this same pattern is observed. Although more advanced agents begin defecting earlier, the initial moves remain the same. The result of this behavior is that future iterations of agents will choose initially matching strategies, instituting some degree of coordination. The benefits of this outcome are obvious. Instead of several turns of enforced waiting as with the non-communicating agents, the agents are able to test their opponent immediately upon commencing the game, and so separate with a high degree of confidence the reciprocating agents from the non-responsive ones. In addition, as the number of strategies remains small, each new policy will not have to change much and the coordinated outcome is enforced. Note that this outcome is not explicitly devised, as in other signaling games where coordination is obviously desirable. Instead, it arises from the decentralized discovery that communication can work to bring players to a better space.

The experimental results confirm that the communicating agents following this strategy achieve higher scores over the total run of the game. Communication is also better for the scores when high reasoning ability is present, both within and between populations. Figure 6 shows the increase in performance for communicating agents as τ rises. As τ increases, the scores begin to diverge from those non-communicating games until the gap between them is over 2.5 points. The performance peaks at a higher reasoning level, and degrades more slowly as reasoning rises. This amount understates the true advantage in scores, because in reality the variance across strategies is much less than the maximum of 24. One of the main reasons for an improvement with communication is that the higher k -step non-communicating agents spend time waiting, which hurts performance against the exploitable members of the population. Communicators have the ability to side-step the waiting issue by directly stating what they are planning to do.

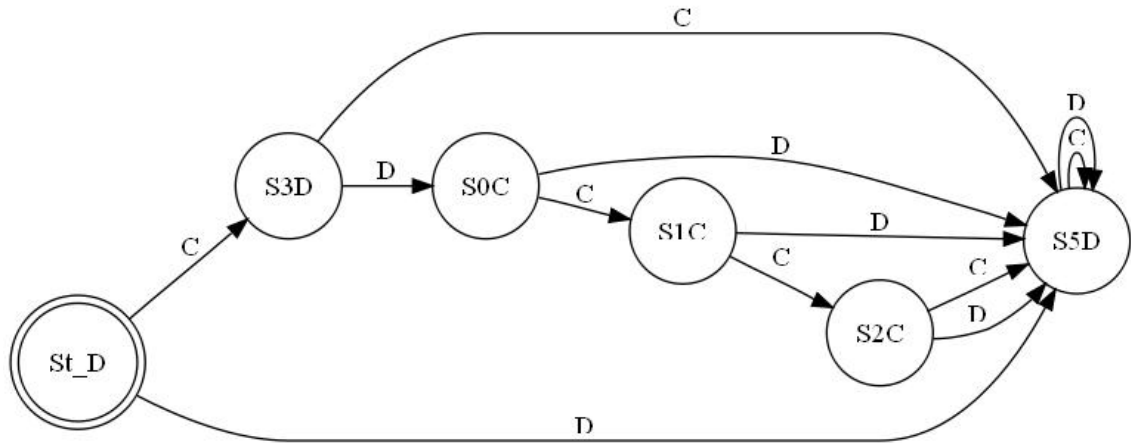


Figure 2: An agent in the six-turn IPD that does not wait before test defecting. $\tau \leq 2$, $k = 1$.

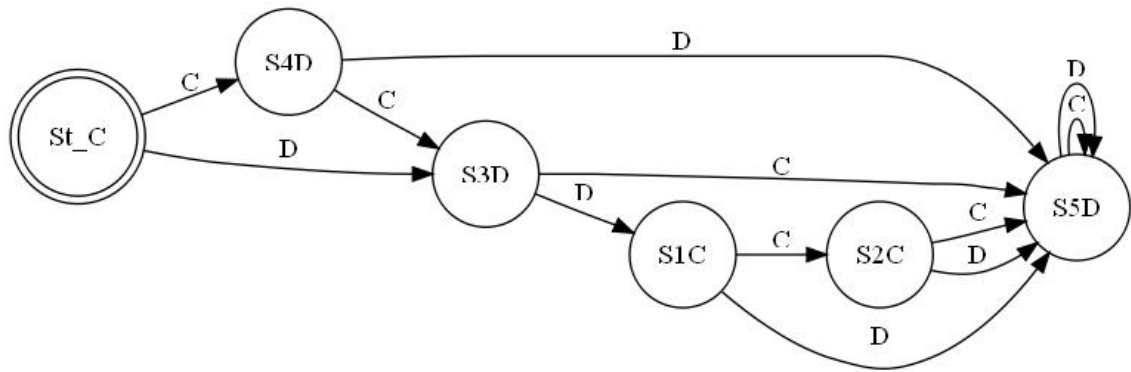


Figure 3: An agent in the six-turn IPD that waits a single turn before test defecting. $\tau \leq 2$, $k = 2$.

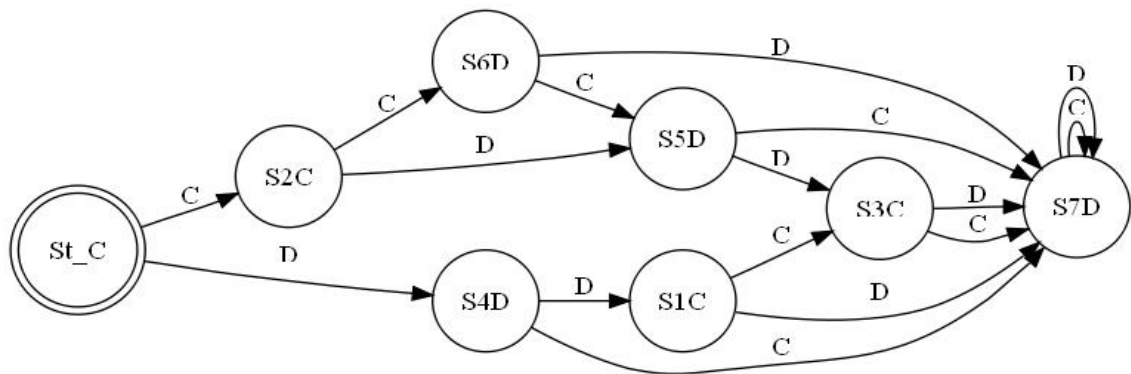


Figure 4: An agent in the six-turn IPD that waits two turns before test defecting. $\tau \leq 2$, $k = 3$.

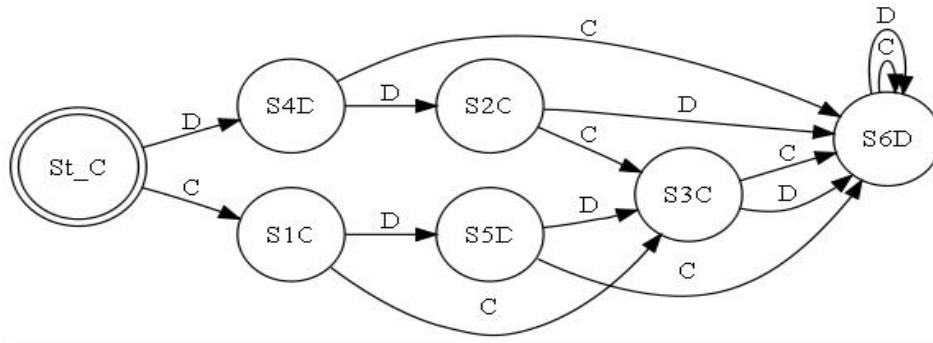


Figure 5: An agent in the six-turn IPD that waits three turns before defecting. Notice that at this stage, it no longer resumes cooperation after these three turns have passed. $\tau \leq 2$, $k = 4$.

5. CONCLUSION

In environments with possibilities for cooperation and conflict, the cognitive hierarchy model provides an approach to planning that combines background expectations about other agents with limited strategic reasoning. In this paper, we have used POMDP inference to explore the strategies derived by the cognitive hierarchy model in limited sessions of repeated prisoners' dilemma.

The solutions show how the cognitive hierarchy model operationalizes the intuitive tradeoffs that characterize interactions in these potentially problematic situations. Agents must challenge their opponents to discover any weaknesses, but they must also engage their opponents to earn their trust. The appropriate balance between these goals, as represented by the round when the agent chooses to test the opponent with a defection, varies as a function both of the distribution of different baseline opponent policies in the model and the extent to which opponents themselves do additional rounds of strategic thinking.

The appropriate balance also depends on the actions available to the agent. In some settings, the cognitive hierarchy model is able to plan truthful announcements of future intentions as a credible way of simultaneously challenging opponents and building trust. Communication thereby improves agents' outcomes.

These preliminary simulations offer an encouraging direction for deeper work on communication and strategy in multi-agent interaction. In future work, we aim both for mathematical results, describing the relationship of baseline strategies, nested inference, and communicative action on planning and performance, and for empirical results, particularly focusing on the extent to which the cognitive hierarchy model can model people's expectations and adaptations and thereby derive agent policies that are able to interact more effectively with users.

Acknowledgments

Thanks to the anonymous reviewers. This research was supported by NSF HSD-0624191.

6. REFERENCES

[1] C. F. Camerer. *Behavioral Game Theory*. Princeton, 2003.

[2] C. F. Camerer, T.-H. Ho, and J.-K. Chong. Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, 104:137–188, 2002.

[3] C. F. Camerer, T.-H. Ho, and J.-K. Chong. A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119:861–898, 2004.

[4] A. R. Cassandra. *Exact and Approximate Algorithms for Partially Observable Markov Decision Problems*. PhD thesis, Department of Computer Science, Brown University, May 1998.

[5] Y. Gal. *Reasoning about Rationality and Beliefs*. PhD thesis, Harvard, 2006.

[6] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

[7] J. M. Keynes. *The General Theory of Employment, Interest, and Money*. 1936.

[8] K. Leyton-Brown and Y. Shoham. *Essentials of Game Theory: A Concise, Multidisciplinary Introduction*. Morgan and Claypool, 2008.

[9] Y. Shoham. Computer science and game theory. *Communications of the ACM*, 51(5), 2008.

[10] J. M. Smith. *Evolution and the Theory of Games*. Cambridge, 1982.