Group Decision and Negotiation, 2:237–258 (1993)

Toward a Theory of Honesty and Trust Among Communicating Autonomous Agents

PIOTR J. GMYTRASIEWICZ
piotr@caen.engin.umich.edu
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109

EDMUND H. DURFEE
durfee@caen.engin.umich.edu
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109

Abstract

This article outlines, through a number of examples, a method that can be used by autonomous agents to decide among potential messages to send to other agents, without having to assume that a message must be truthful and that it must be believed by the hearer. The main idea is that communicative behavior of autonomous agents is guided by the principle of economic rationality, whereby agents transmit messages to increase the effectiveness of interaction measured by their expected utilities. We are using a recursive, decision-theoretic formalism that allows agents to model each other and to infer the impact of a message on its recipient. The recursion can be continued into deeper levels, and agents can model the recipient modeling the sender in an effort to assess the truthfulness of the received message. We show how our method often allows the agents to decide to communicate in spite of the possibility that the messages will not be believed. In certain situations, on the other hand, our method shows that the possibility of the hearer not believing what it hears makes communication useless. Our method thus provides the rudiments of a theory of how honesty and trust could emerge through rational, selfish behavior.

Key words: distributed artificial intelligence, rational communication, multiagent systems, decision making, agent modeling, belief

1. Introduction

Agents that are fully autonomous should be able to freely choose among all actions—physical and communicative—available to them. The challenge in designing such systems, though, is to develop the theories and methods needed by the autonomous systems to use their capabilities rationally. In this article we extend the method presented in Gmytrasiewicz, Durfee, and Wehe (1991a) to allow an agent to choose rationally what to communicate to another agent, without the

This research was supported, in part, by the Department of Energy under contract DG-FG-86NE37969, and by the National Science Foundation under grant IRI-9015423.

assumption that the transmitted information has to be truthful, and is guaranteed to be believed. The communicating agent, therefore, has to consider that blatant lies could be easily detected and therefore are not likely to be believed. Sending such messages thus could make no sense. In fact, sometimes truthful messages might be unbelievable, and sending them could be irrational, too. On the other hand, some lies might be believable, or at least could force the hearer to change its actions regardless of whether it believes them or not. The goal of the speaking agent is to choose to send messages that will impact the hearer to the speaker's advantage. Assessing whether messages, true or false, will be believed is thus important to the speaker.

A fundamental paradigm underlying our approach is that autonomous intelligent agents use the principle of maximization of their subjective expected utility in all of their purposeful undertakings. This paradigm is a central one in utility theory, decision theory, and game theory, on which we draw in our work. Our adopting the utilitarian paradigm absolves us from considering more specific issues: for instance, whether robots should lie, when, or whether they should be gullible. Our aim is to ensure that they be rational whatever they do. If lying is rational, so be it.

What we find interesting is to understand why rational, selfish agents would ever choose to tell the truth, believe in what they hear, or even communicate at all, without the external imposition of a protocol (Zlotkin and Rosenschein 1991) or incentives (Ephrati and Rosenschein 1991) that force them to be honest. Clearly, among people, honesty and trust can emerge among friends, while deception and disbelief often predominate among adversaries. Our motivation is to understand how rational decisions among these options are made, and possibly what their social implications are (see section 6). Our study also has practical importance for designing autonomous agents that can function in adversarial situations, such as competitive markets and battlefields.

The issue of lies in communication has recently been addressed in the Distributed AI literature by Zlotkin and Rosenschein (1989, 1990, 1991). In their work, Zlotkin and Rosenchein analyze the use of lies in negotiation and conflict resolution in various domains. They study how preestablished negotiation protocols, for particular domains, can ensure that agents will not lie. We, on the other hand, do not assume a prearranged protocol. They also suggest a taxonomy that divides lies into false information about the broadcasting agent itself or false information about the sender's future actions. This distinction is essentially identical to our examination of lies in modeling and intentional messages in this article. Unlike our work, however, they do not consider the reciprocal issue of whether a message recipient should believe what it hears.

The complications created by the possibility of dishonest communication have also been analyzed in the economics literature (Mayerson 1988). There, issues of designing a communication system that optimizes the information exchanged by rational agents are discussed. A number of these issues are identical to the ones we are dealing with, but the emphasis is on the communication channel design,

as opposed to the decision making of the individual agents motivated by maximization of the expected utility of the messages exchanged.

Our approach is based on the Recursive Modeling Method (RMM) (Gmytrasiewicz, Durfee, and Wehe 1991b), and on our analysis of how communication transforms the RMM hierarchy (Gmytrasiewicz, Durfee, and Wehe 1991a) (outlined also in the next section), which is intended to be a complete representation of an agent's knowledge relevant to the decision-making process in a multiagent environment. As in our earlier work, the main guideline to solve the problem is the recursive use of the intentionality assumption, that is, the assumption that other agents are rational and seek to maximize their expected utility. To evaluate the utility of a message, the sender of the message will attempt to predict if the receiver will believe it or not. The receiver, on the other hand, will attempt to guess if the sender was transmitting the truth. The intentionality assumption applied by the receiver to the sender will then be used to answer the question of whether it would pay for the sender to lie. We get here a recursive pattern on the communicative level, and the resulting recursive hierarchy, which we call a communication hierarchy, can be solved by methods similar to those used in the case of recursive hierarchies containing physical actions, called action hierarchies and analyzed in Gmytrasiewicz, Durfee, and Wehe (1991a, 1991b). However, as it turns out, each of the recursive levels of communicative options of the agents requires solution of at least one action hierarchy, as opposed to the solution of a single payoff matrix in action hierarchies. In addition, solving a communication hierarchy introduces the notion of playing against "nature" as we shall see, which fundamentally changes the analysis such that some forms of ambiguity—whether the hearer should believe the speaker, for example—might be impossible to resolve definitively.

As in our previous work, the evaluation of the utility of a message, M, will be based on the following equation:

$$U(M) = U_{pM}(Y) - U_p(X). \tag{1}$$

This equation expresses the utility of a message, M, as the difference between the expected utility, $U_p(X)$, of the best action, X, before sending the message, and the expected utility, $U_{pM}(Y)$, of the preferred action, Y, after the message was sent.

In the remainder of the article, we first illustrate the basic concepts and complications introduced by the possibility of the hearer not believing a message, using a simple and intuitive example (section 2). Then, we consider the utility of lying in messages that describe the environment, which we call modeling messages, using an example of a lie that pays off, and one that does not (section 3). We go on to discuss the issues of lies and belief in messages describing the intentions of the speaker, called intentional messages (section 4). We show how the possibility of the messages being dishonest and not believed can lead to the agents not engaging in communication at all, using the example of the Prisoner's Dilemma (section 5). We do not consider the issues of lying and belief in the cases

of acknowledging messages, questions, and imperatives in this paper. The difficulties introduced by communication channel unreliability are also ignored for simplicity.

2. A simple example

In this example we will consider a simple modeling message that can be exchanged between two agents. As we mentioned, modeling messages contain information about the environment in which the interaction takes place or about the properties of the agents involved. Let us consider the example of the two interacting agents with two goals in the environment, valuable to both of the agents, as depicted in Figure 1. We will assume in the following analysis that the agents cannot see through walls, which is common knowledge.

Let us consider agent R1 contemplating the value of the following modeling message M1: "There is G1′ behind the wall," intended for agent R2. Our analysis will use a number of payoff matrices. The payoffs in these matrices are computed as a sum of the values of all of the performed goals minus the cost personally incurred in the process by the individual agents. For example, if R1 pursues G1′ and R2 pursues G1, then the payoff for R1 is (2 + 2) − 2 = 2. The labels G1 and G1′ of rows and columns stand for the goals G1 and G1′ the agents may pursue, while S stands for the option "stay still or do something else." To make our example concrete for the reader interested in the quantitative calculations, we represent the views of interactions in terms of payoff matrices in this article. Readers interested in a qualitative understanding need not be concerned with the specific values in these matrices.

$P^{R1}_{G1'}$ is the payoff matrix describing R1's decision-making situation:

		R2		
		G1	G1′	S
	G1	1	3	1
R1	G1′	2	0	0
	S	2	2	0

Let P^{R2} be R2's payoff matrix describing the situation in which R2 does not know about the goal G1′:

		R1	
		G1	S
R2	G1	0	0
	S	2	0

Note that the P^{R2} matrix also describes the case in which R2 has received the message M1 but decided not to believe it, since, if R2 does not believe M1, it still does not believe that G1′ exists.

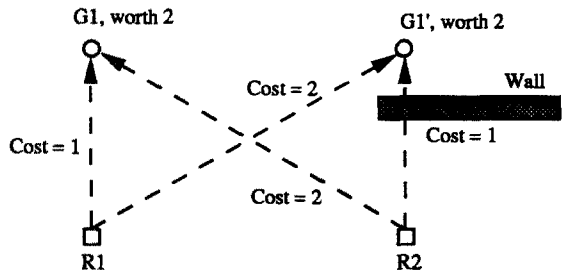

Figure 1. Example scenario.

Let P^{R1} be the matrix of R1's payoffs if it were the case that G1′ were not behind the wall:

		R2	
		G1	S
R1	G1	1	1
	S	2	0

Finally, let us denote by $P^{R2}_{G1'}$, the payoff matrix describing R2 knowing about G1′:

		R1		
		G1	G1′	S
	G1	1	3	1
R2	G1′	2	0	0
	S	2	2	0

With these matrices, R1 can build the recursive hierarchy (Gmytrasiewicz, Durfee, and Wehe 1991b) before communication takes place, shown in Figure 2. R1 knows that R2 does not see G1′ and models it on the second level using the matrix P^{R2}. R1, thinking that R2 is unaware of G1′, models R2's model of R1's decision making without G1′, that is, using the matrix P^{R1} on the third level. The rest of the levels are constructed in a similar way. Thus, the goal G1′ is not present at all on the lower levels of the hierarchy in Figure 2. Agent R1 can easily solve this hierarchy (Gmytrasiewicz, Durfee, and Wehe 1991b) by propagating the information in the hierarchy bottom-up (starting from any level below the second in this case). Let us say that we use the hierarchy ending with the matrix P^{R1}. If we look no further down, then R1 should consider either of R2's moves as equally likely, so R1 should expect equivalent average payoffs for either of its moves. Propagating this conclusion up the hierarchy, R2 would consider either of R1's moves as equally likely in the matrix P^{R2}, and will thus prefer S. Moving up another level, now R1 would expect R2 to take move S, so R1 should prefer move G1, which, going up the hierarchy will again cause R2 to prefer S. The choices of

R1 preferring G1 and R2 preferring S progress all the way to the level right below the top, and R1 should thus conclude that R2 will pursue its option S, i.e., R2 will stay put. The best option of R1 then is to pursue G1 with its payoff of 1.

2.1. First recursive level—R1's view of its own knowledge

Let us now see what the state of R1's knowledge would be if it were to transmit message M1 to R2, stating: "There is G1′ behind the wall." At this point we relax the assumption in our previous work that messages are always truthful and always believed. R1 is, therefore, uncertain whether R2 will believe M1. This uncertainty is reflected as a branching of the hierarchy, depicted in Figure 3, at the top. The

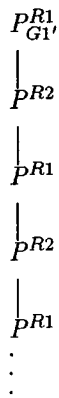

Figure 2. Recursive hierarchy before communication for simple example.

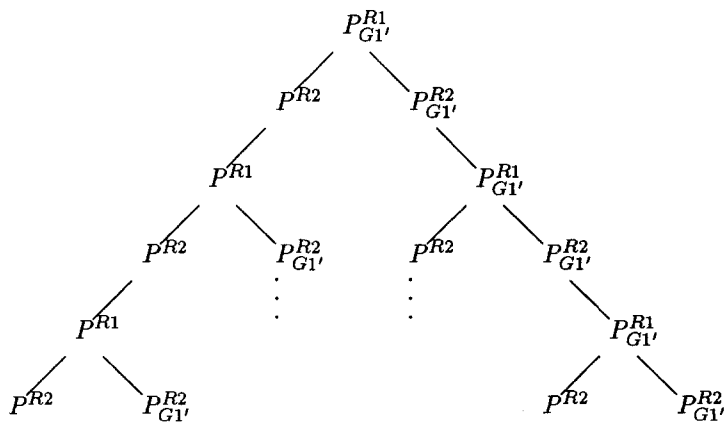

Figure 3. State of R1's knowledge due to sending M1.

right branch corresponds to the possibility of R2 believing M1, and the left branch corresponds to the possibility of R2 not believing it. On the level below, R2, if it believed M1, will model R1 as having G1′ in its world model. If R2 did not believe M2, it will model R1 as not believing it, too, of course. The branching below the third level corresponds to R2 being aware that R1 does not know whether M1 was believed or not, in either case of it being true or not.

At this stage, R1 does not have any information on whether R2 is going to believe M1. Consequently, the uncertainties illustrated by branching in the above hierarchy can be treated equiprobabilistically.[1] The result of solving the hierarchy is the conclusion that R2 will pursue G1′ if it believed M1, and stay still if it did not. Thus, the probabilities of R2's pursuing G1, G1′, and S, can be summarized by the following intentional probability distribution: $P^{R1}_{R2} = (0,0.5,0.5)$. The best option of R1, then, is to pursue G1, and the expected utilities for R1 are 3 and 1, for the cases of R2 believing M1 or not, respectively. These values can be used to compute the value of M1 (recall that without communication R1 would get 1) as being 2 and 0 in each of these cases, respectively. They can be summarized in the following communicative matrix, which we will call $CV^{R1}_{G1'}$:

	R2	
	B	not-B
$V^{R1}(M1)$	2	0

The options of agent R2, B and not-B, correspond to R2's believing M1 or not, respectively. Since, at this stage, R1 does not know whether R2 will believe M1, it would treat these possibilities as equiprobable and evaluate that the message M1 would improve the quality of the interaction and raise R1's expected utility by 1.

2.2. Second recursive level—R1's view of R2's reasoning about M1

As we just derived, the actions of R2 in the case of it believing M1 or not are to pursue G1′ or S, respectively. But will R2 decide to believe M1? R2's decision making will again depend on the expected payoffs it gets in each case of M1 being true or not, and the likelihood that M1 is true.[2] These payoffs depend on what R1 would do in each of these cases. If M1 is true, R1's action can be predicted from the hierarchy in Figure 3, which results in R1's best action being G1. If M1 is false, on the other hand, R1's action can be predicted from the hierarchy in Figure 4.

It is a hierarchy very similar to the one in Figure 3, with the matrix P^{R1} on the top instead $P^{R1}_{G1'}$. This hierarchy results in the best option of R1 being G1.

Thus, R1 can see R2 as having two options: it can believe M1 or not. In either case of M1 being true or not, R2 knows that R1 will go G1. If R2 believes M1, it would pursue G1′ getting 3 if G1′ is there, and getting only 1 if G1′ is not there.

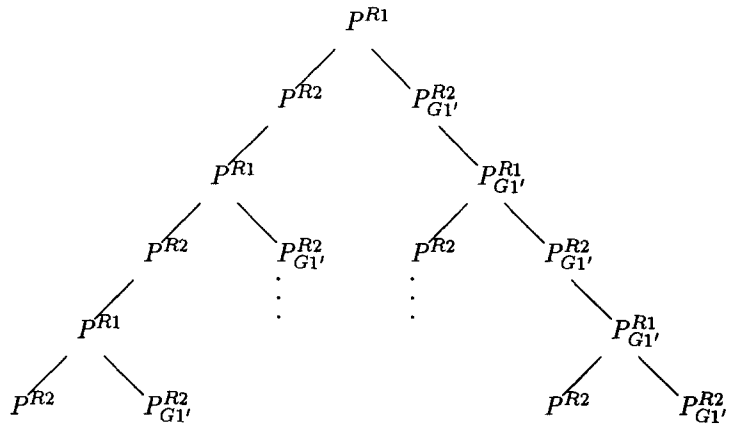

Figure 4. State of R1's knowledge after sending M1, if M1 went not true.

If R2 does not believe M1, it would stay still and get a payoff of 2 whether G1′ is behind the wall or not. Let us assemble these results into the following communicative matrix, which we will call C^{R2}:

		Nature	
		M1-T	M1-F
R2	B	3	1
	not-B	2	2

Therefore, the decision of R2 as to whether to believe M1 depends in this case on R2's judgment on whether M1 is actually true or not. Something that should be stressed about this matrix is that it represents R2's options of believing M1 or not in rows, but the columns do not represent R1's options as in the cases considered in Gmytrasiewicz, Durfee, and Wehe (1991a, 1991b). The columns represent the state of *Nature* instead. This is because it is not in R1's power to make M1 true or false.

To summarize the above analysis, R1, in its evaluation of the value of M1 on the first level, sees that this value depends on whether R2 will believe it or not, which in turn depends on whether R2 thinks M1 is true or not—based on the second level. R2 may attempt to answer this question by reasoning about R1 in each of these possible worlds. These considerations belong to the third recursive level.

2.3. Third recursive level—R1's view of R2's view of R1

On this level, R1 goes deeper and models R2's attempt to answer the following question: Am I in the world in which M1, just received, is true, or am I in the world in which M1 is false? In answering this question, R2 can use the intention-

ality assumption about R1, model R1 in each of these worlds, and determine which of these worlds is consistent with R1's action of transmitting M1. Let us here digress from the example and point out a limitation of our approach. It may happen in some cases that the hearer's attempt to judge the truthfulness of a message based on the fact that it was transmitted by a rational speaker does not lead to conclusive results because the action of a rational speaker may be consistent both with the world in which the message is true, and the world in which it is false. The methods that can be employed in these cases would have to rely on prior knowledge or the use of a suitably chosen test based on which the hearer can decide on the truthfulness of what it hears.

Returning to the example, R2 sees R1 as having the options of transmitting M1 or not. In the world in which M1 is true, R1's transmitting M1 would give it a payoff of 3 if R2 believes, or payoff of 1 if R2 does not believe, as computed before. Let us assemble these values into a matrix $C^{R1}_{G1'}$:

		R2	
		B	not-B
R1	send M1	3	1
	not-send M1	1	1

The above matrix closely corresponds to the matrix $CV^{R1}_{G1'}$ in which the expected utility of the message itself was summarized. The matrix $CV^{R1}_{G1'}$, can be obtained from $C^{R1}_{G1'}$ with the help of equation (1) relating the utilities of actions with and without communication to the value of this communication.

In the world in which M1 is false, R1 can expect payoffs assembled in the following matrix C^{R1}:

		R2	
		B	not-B
R1	send M1	1	1
	not-send M1	1	1

From the communicative matrices C, a recursive communication hierarchy, depicted in Figure 5, may be assembled.

This hierarchy fully illustrates the recursive nesting of beliefs of agent R1 and is similar to the action hierarchies considered in Gmytrasiewicz, Durfee, and Wehe (1991a, 1991b). The important difference lies in its analysis and method of solution. Let us note that, as mentioned before, the rows of matrices $C^{R1}_{G1'}$ and C^{R1} on the third level cannot be directly translated to the columns of the matrix C^{R2}. This is due to the fact that R2's opponent in C^{R2} is Nature, not R1, who is a player in both $C^{R1}_{G1'}$ and C^{R1}. Thus, the above hierarchy cannot be solved by directly propagating conclusions of the third level to the second level (and similarly, from fifth to fourth, and so on). While direct propagation is not possible, R2 can analyze the matrices $C^{R1}_{G1'}$ and C^{R1}, note that the value of the message M1 it just heard according to C^{R1} is zero, and thus conclude that only the world in which M1 is true, one

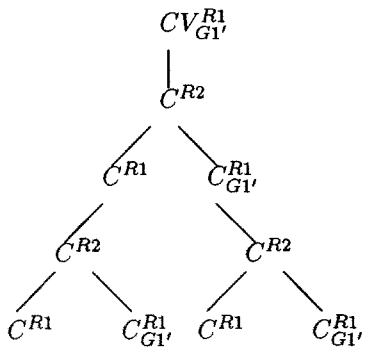

Figure 5. Recursive communication hierarchy.

in which R1 is modeled by $C^{R1}_{G1'}$, is consistent with R1 sending the message M1. In other words, R2 could see that if G1′ were not there, it would not pay for R1 to lie and say that G1′ is there. This observation can be used in the matrix C^{R2} to conclude that R2 will believe M1, and, as a final conclusion, that the value of message M1 to R1 is equal to 2.

The above example was quite intuitive. R1 was contemplating a true message, found the message's impact on R2, and was able to determine that R2 will believe this message despite risks of doing so. That was due to the character of the interaction the agents were engaged in; it would simply not make any sense for R1 to lie.

3. Lying in modeling messages

In this section, we build on the intuitions developed in the previous section to analyze two examples of lies in modeling messages.

3.1. A lie that pays

Let us consider a scenario similar to the "phantom letters" scenario considered in Zlotkin and Rosenschein (1990) and depicted in Figure 6. Let us say that R1 contemplates sending a message, M2, to R2 stating, "There is a G2, worth 15, at Location 1." Let us define the following payoff matrices. P^{R1} will describe R1's payoffs:

		R2	
		G1	S
R1	G1	4	4
	S	10	0

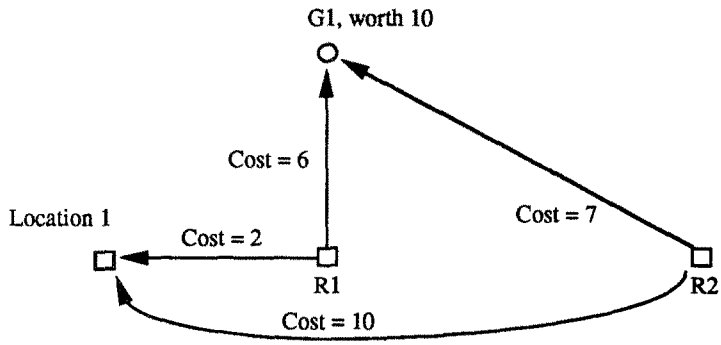

Figure 6. A phantom goal scenario.

P^{R2} is R2's payoff matrix without G2 there, and if R2 did not believe M2:

		R1	
		G1	S
R2	G1	3	3
	S	10	0

Let P^{R1}_{G2} summarize R1's payoffs if G2 were really at Location 1:

		R2		
		G1	G2	S
	G1	4	19	4
R1	G2	23	13	13
	S	10	15	0

P^{R2}_{G2} will then contain R2's payoffs if it believed M2:

		R1		
		G1	G2	S
	G1	3	18	3
R2	G2	15	5	5
	S	10	15	0

R1's state of knowledge before M2 is sent can be represented by the hierarchy in Figure 7.

This hierarchy does not converge directly; it flip-flops between R2's pursuing G1 or not as progressively deeper levels are considered, which can be summarized in the following intentional probability distribution: p^{R1}_{R2} = (0.5,0.5). The best option of R1 is then to stay put and expect the payoff of 5.[3]

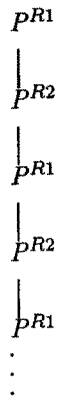

Figure 7. Recursive hierarchy before communication for a lie that pays.

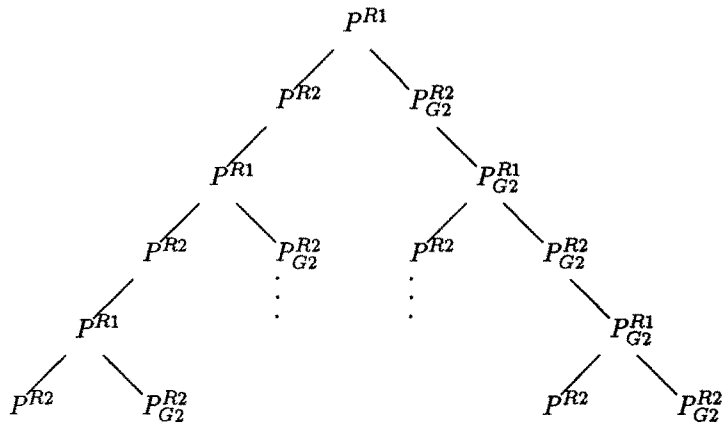

Figure 8. Recursive hierarchy of R1 if M2 were transmitted.

3.1.1. First level of recursion. On this level, R1 is reasoning about its knowledge state, depicted in Figure 8, if it were to transmit M2.

In the process of analyzing this hierarchy, R1 can arrive at the unique estimate of R2's behavior: $p^{R1}_{R2} = (1,0)$. Thus, R1 estimates at this level that R2 will pursue goal G1 in either case of it believing M2 or not. That leaves doing nothing as R1's best option, as before sending M2, but now with its expected payoff of 10! R1's reasoning on the first level, therefore, results in an estimate of M2's utility of 5. The communicative matrix, CV^{R1}, summarizes these results:

	R2	
	B	not-B
$V^{R1}(M3)$	5	5

*3.1.2. **Second level.*** At the second level, R1 can reason about R2's options to believe M2 or not, and the payoffs it expects in each case. R2's action if it believes M2 can be computed based on the right subtree of the hierarchy in Figure 8. Its analysis results in the best option of R2 being the pursuit of G1. If R2 does not believe M2, its knowledge is represented by the left subtree in Figure 8. Its analysis reveals that the best option R2 has in this case is also to pursue G1. This somewhat surprising conclusion arises since R2 knows that R1 will stay still after transmitting M2, even if it does not know whether M2 was believed. Thus, R2 is best off pursuing G1 even if it thinks that M2 is a lie.[4] If M2 is true, R2 would expect R1 to pursue G2, and its own pursuit of G1 would result in the total payoff of 18. If M2 is false, R2 will get only 3. These values can be summarized in communicative matrix C^{R2}:

		Nature	
		M2-T	M2-F
R2	B	18	3
	not-B	18	3

As the above matrix shows, it really does not matter for R2 whether it should decide to believe M2 or not, since both options offer the same payoffs. Since its choices do not matter, R2 does not have any incentive to extend its analysis to deeper levels (other than to possibly find out if R1 is a liar or not, which does not matter for this particular case).

*3.2.3. **Third level.*** If R2 were to engage in considerations on the third recursive level, it would discover that it pays for R1 to transmit M2 both in the case when it is a lie and when it is true. The communicative payoff matrices it would then construct for R1 corresponding to the cases of M2 being true and being false, respectively, are depicted below.

C^{R1}_{G2} is:

		R2	
		B	not-B
R1	send M2	23	13
	not-send M2	13	13

C^{R1} is:

		R2	
		B	not-B
R1	send M2	10	10
	not-send M2	5	5

3.2. A lie that does not pay

In this subsection, we summarize an example [see Gmytrasiewicz and Durfee (1992) for details] very similar to the one considered in the previous subsection, with a slightly different assignment of values to the goals and their costs, as depicted in Figure 9. We assume that R1 contemplates sending a message, M3, to R2 stating, "There is a G3, worth 4, at Location 1."

The four payoff matrices describing the decision-making situations of agent R1 in the cases of M3 being true or not, and of agent R2 if it believed M3 or not, can be constructed just as in the preceding subsection. Also, the hierarchies of the payoff matrices can be built analogously.

The analysis of the hierarchy that describes the state of R1's knowledge before sending M3 reveals that R1's best option in this case is to pursue G1 with the payoff of 2. It also turns out that R1, engaged in reasoning at the first level, would conclude that even if it were to send M3, pursuing goal G1 would remain its best option with its payoff of 2, and thus sending M3 does not pay off. The result of the analysis on the second level, during which R1 considers R2's options of believing M3 or not, is that M3 would not be believed. This conclusion changes when the third level is entered, since it would become clear for R2 that sending M3 does not pay off for R1 in the case when G3 is not at Location 1. Consequently, the reasoning on the third level leads to the conclusion that M3 would be believed. The fourth level reverses this conclusion again, and a pattern becomes apparent according to which the analysis of the recursive communicative hierarchy flip-flops from level to level. As we mentioned before, our suggestion of a way out of this impasse coincides with one described in Gmytrasiewicz, Durfee, and Wehe (1991b) for the case of action hierarchies. The recursive analysis simply does not provide a conclusive answer in this case, beyond telling R1 that R2 will either believe M3 or not. Treating these possibilities as equally likely, R1 would finally conclude that it is the best for it to pursue G1, and that transmitting M3 does not pay off.

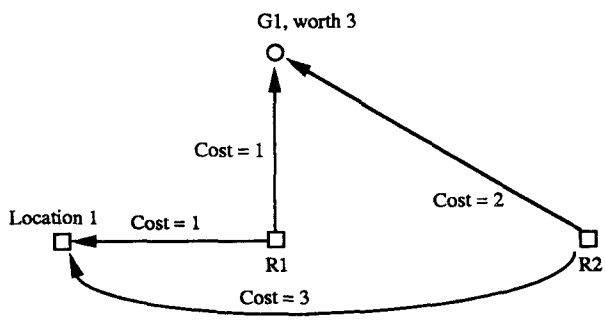

Figure 9. A variation of phantom goal scenario.

3.3. Discussion

The examples above, of the lie that pays and of the lie that does not pay, show how the parameters of the particular interaction the agents are involved in impact their communicative decisions. In the first example, the hearer, R2, is left virtually without a choice; it must pursue the goal it knows about no matter whether it really believes in the phantom goal or not. In the second example, on the other hand, R1 could not expect R2 to believe the phantom goal message and concluded that it does not pay to lie to R2.

Let us also point out another interesting feature of the first example of a lie that pays (for R1). Note that, without communication, if both agents used the RMM algorithm, they would both conclude that staying still is better than pursuing goal G1. In this situation, goal G1 would be left unattended altogether, while both agents are being rational. This situation is similar to one arising in the Prisoner's Dilemma game, in which two rational players choose a defecting move and are left with a meager payoff (section 5). Also note that if both agents decided to pursue G1 without communication, the cumulative cost would outweigh the value of achieving G1.

4. Lying and belief in intentional messages

An intentional message expresses a commitment, by the agent that transmits the message, to an action.[5] Apart from the fact that an intentional message impacts the speaker and the hearer differently than a modeling message, our analysis here will be very similar to the case of modeling messages. The speaker will attempt to assess the value of a given intentional message, which in general will depend on whether the hearer believes it. The hearer, on the other hand, will attempt to guess the truth value of the received message by wondering if it pays for the speaker to stick to its expressed commitment or not.

Let us first consider a scenario depicted in Figure 10.

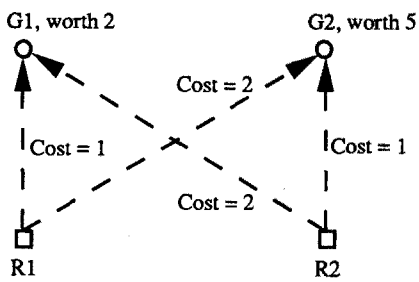

Figure 10. Example of interacting agents.

R1's payoff matrix will be called P^{R1}:

		R2		
		G1	G2	S
	G1	1	6	1
R1	G2	5	3	3
	S	2	5	0

P^{R2} will contain R2's payoffs:

		R1		
		G1	G2	S
	G1	0	5	0
R2	G2	6	4	4
	S	2	5	0

The recursive hierarchy containing R1's knowledge before communication takes place is depicted in Figure 11.

The solution to the above hierarchy for R2's intentional probability distribution is: $p^{R1}_{R2} = (0.25,0.5,0.25)$, which results in the expected utilities of R1's options G1 and G2 being both 3.5. R1's option S has expected utility of 3, so R1 cannot hope for more than 3.5 without communication. Let us see how R1 can better its utility by sending a truthful intentional message, M1, saying "I will pursue G1."

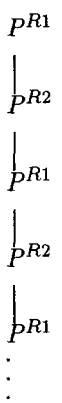

Figure 11. Recursive hierarchy before communication.

4.1. First recursive level

At this level R1 will analyze the state of its knowledge resulting from sending M1, and attempt to assess M1's utility. Let us introduce a payoff matrix, P^{R2}_{M1}, of R2, containing its options and their payoffs if it believed M1:

		R1
		G1
	G1	0
R2	G2	6
	S	2

The recursive hierarchy of R1 after sending M1 is depicted in Figure 12. The right branch corresponds to the possibility that the hearer believes the message and thus does not have to engage in modeling the speaker at a deeper level to predict its intentions.

The result of solving this hierarchy for the intentional probability distribution of R2 is: $p^{R1}_{R2} = (0,1,0)$, which means that R1's best option is G1 with its expected utility of 6. Thus, if R2 believes M1, it will pursue G2, but even if it does not believe M1, it will still pursue G2. The summary of R1's assessment of the value of M1 at this stage will be called CV^{R1}:

	R2	
	B	not-B
$V^{R1}(M1)$	2.5	2.5

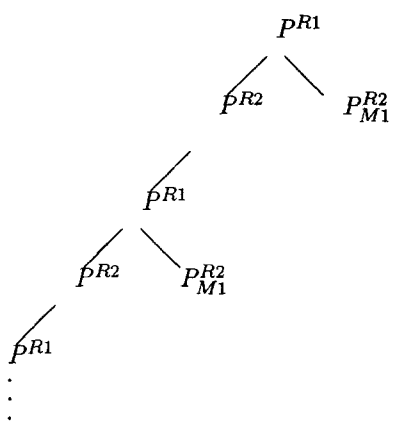

Figure 12. Recursive hierarchy of R1 after M1 was sent.

At this point, R1 loses interest in whether R2 will believe M1 or not since R1's payoffs do not depend on these possibilities. The value of M1 to R1 can be established at 2.5. If, for academic purposes, R1 did consider R2's decision making about whether to believe M1 or not, it would notice that R2 does not care about M1's truthfulness either.

5. Communication in Prisoner's Dilemma game

We will now turn to the issue of agents engaged in the game of Prisoner's Dilemma (PD) considering communicating their intentions. Prisoner's Dilemma is an abstraction of many different real-life situations, ranging from trench warfare to evolutionary genetics (Axelrod 1984). Traditionally, game theorists do not analyze the PD game in the context of the players communicating because it is considered that communication in this game is either useless or dishonest. Our aim in this section is to show that these assumptions can be derived using our approach.

Let us call the payoff matrix of the player I in the PD matrix P^I:

		II	
		c	d
I	C	3	0
	D	5	1

P^{II} will be the corresponding matrix of player II:

		I	
		C	D
II	c	3	0
	d	5	1

Further, let us define P^{II}_{I-C} as:

		I
		C
II	c	3
	d	5

and P^{II}_{I-D} as:

		I
		D
II	c	0
	d	1

The rational choice of a move in a one-shot play of PD without communication is to move D for player I, and d for player II. Let us say that player I considers sending an intentional message, called USELESS, that states "I will choose D." The state of knowledge of player I can then be depicted as the hierarchy in Figure 13.

After analyzing this hierarchy, player I has to conclude that $P^I_{II} = (0.1)$, i.e., player II will choose d whether it believes USELESS or not. That means that the best option of player I is D, with its payoff of 1, and the value of USELESS is 0. Going down the recursive levels does not make the value of USELESS any dif-

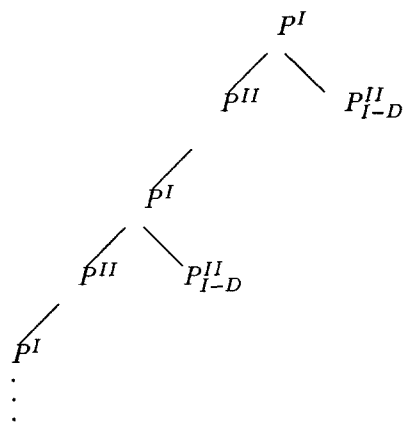

Figure 13. Recursive hierarchy of player I after USELESS was sent.

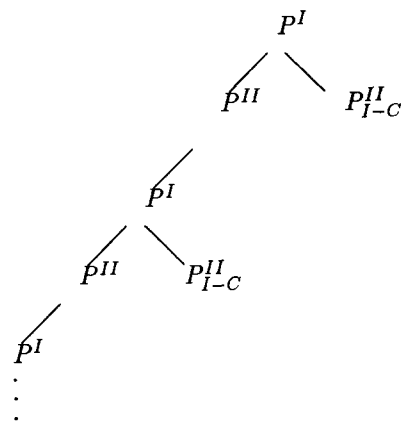

Figure 14. Recursive hierarchy of player I after DISHONEST was sent.

ferent than 0, since there is no way that player II would be convinced by USELESS to choose c.

Let us instead, then, consider another intentional message that player I may consider. It would state "I will choose C." Let us call this message DISHONEST. Player I's state of knowledge on the first recursive level is included in figure 14.

Again, we have: $p^I_{II} = (0,1)$, i.e., player II chooses his move D whether it believed DISHONEST or not. Given that fact, D is player I's best choice, and, as rational, player I would choose it, which earns DISHONEST its name. Were player I to go deeper into the recursive levels, he would conclude that player II will never believe DISHONEST, of course.

The two messages above exhaust player I's choices of intentional messages, and there are no other messages of importance it can consider further. Player I will therefore conclude that it does not pay to talk to player II at all.

6. Discussion and future work

Allowing the possibility that messages can be lies and disbelieved adds a substantial amount of overhead to the computation of expected values of candidate messages. The analysis of each of the recursive levels now requires the solution of at least one newly created action hierarchy. Furthermore, fairly complex reasoning was often necessary to propagate the results of the recursive levels upward. This complexity would make the routines that rigorously employ the ideas presented here more costly.

The examples considered in this article were intended to illustrate both the power and the limitations of our approach. For the senders of potential messages, it was always possible to assess the value of both modeling and intentional messages. On the hearer's side, however, we showed that our approach may not suffice to make a decision on whether to believe the incoming information for modeling messages in certain situations. These circumstances may render the recursively applied intentionality assumption useless, and other means may have to be employed.

In some cases, the messages that we analyzed turned out to be not worth sending. That points to the possibility that the agents will not find any suitable messages to exchange at all, as we showed in the case of the Prisoner's Dilemma. That simply means that interactions between agents in these situations will be silent.

We find it interesting to consider the results of this article in the context of the following example. Let us imagine a rich and dynamic environment populated by autonomous, heterogeneous agents. We think that the insights gained in this article can be used to predict that some agents will turn out to converse a great deal, while not communicating with the others. The fact that communication is likely to be repetitive brings about questions, already explored in Gmytrasiewicz, Durfee, and Wehe (1991b) for decision making about physical actions, about how

repetition will influence the whole idea of lies and belief. We think that it is fairly safe to qualitatively extend the results obtained in Gmytrasiewicz, Durfee, and Wehe (1991b) to these issues. Thus, just as repetitive interactions facilitates cooperation of the agents in their choice of physical action, so, too, will it promote cooperative traits in communication. That means that agents who communicate repetitively will tend to tell the truth to each other and to trust each other, while "outsiders" will likely be lied to and not believed.

Our future research will attempt to formalize more fully the process of analyzing communicative payoff hierarchies. We also hope to apply these formal methods to the cases of repeated communication which should allow us to perform exhaustive testing of the hypothesis that repetitive communication would allow honesty and trust to spontaneously arise in societies of selfish, rational agents.

Notes

1. We use here the principle of indifference to represent the lack of knowledge by a uniform probability distribution, which contains no information. That, of course, does not preclude the possibility that other knowledge might be available. If it were, it would be represented as a probability distribution and used here instead.
2. We discount here the possibility that the hearer might choose to believe a message because it would be good if it were true as irrational. Of course, acting as if it were true based on the expected utility of doing so can be rational, as we will see.
3. As we remark in Gmytrasiewicz, Durfee, and Wehe (1991b), the inability of the analysis to converge on a unique solution simply means that the information contained in the hierarchy is inconclusive. While a conclusive answer is obviously desirable, we must avoid the temptation to have the system imply that it can conclude more than it should. Thus, rather than jump to a conclusion, and perhaps use ad hoc methods to program agents to jump to the same conclusion, we prefer to explicitly represent and use the uncertainty about conclusions.
4. This situation is similar to the following everyday occurrence. Both you and your friend want a certain thing done, but it seems that it would be somewhat more convenient for him to do it. He knows that, and you know that he knows, but at a certain point he says that he has something else, much more important, to do. Now, you think it is just an excuse and you do not really believe him, but the bottom line is that he will not do what you had previously assumed. Thus, the only option you have is to do it yourself, even if you thought that the excuse was a lie.
5. The intention can also be probabilistic. An agent can, for instance, express its intention to pursue two of its options with probabilities 0.3 and 0.7.

References

Axelrod, Robert. (1984). *The Evolution of Cooperation.* New York: Basic Books.

Ephrati, E., and J.S. Rosenschein. (1991). "The Clarke Tax as a Consensus Mechanism Among Automated Agents." In *Proceedings of the National Conference on Artificial Intelligence,* San Jose, CA, July, AAAI Press, pp. 173–178.

Gmytrasiewicz, Piotr J., and Edmund H. Durfee. (1992). "Truth, Lies, Belief and Disbelief in Communication Between Autonomous Agents." In *Proceedings of the Eleventh International Workshop on Distributed Artificial Intelligence,* February.

Gmytrasiewicz, Piotr J., Edmund H. Durfee, and David K. Wehe. (1991a). "The Utility of Communication in Coordinating Intelligent Agents." In *Proceedings of the National Conference on Artificial Intelligence*, July, AAAI Press, pp. 166–172.

Gmytrasiewicz, Piotr J., Edmund H. Durfee, and David K. Wehe. (1991b). "A Decision-Theoretic Approach to Coordinating Multiagent Interactions." In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, August, Los Altos, CA: Morgan Kaufmann, pp. 62–68.

Mayerson, Roger B. (1988). "Incentive Constraints and Optimal Communication Systems." In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, March, Los Altos, CA: Morgan Kaufmann, pp. 179–193.

Zlotkin, Gilad, and Jeffrey S. Rosenschein. (1989). "Negotiation and Task Sharing among Autonomous Agents in Cooperative Domains." In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, August, Los Altos, CA: Morgan Kaufmann, pp. 912–917.

Zlotkin, Gilad, and Jeffrey S. Rosenschein. (1990). "Blocks, Lies, and Postal Freight: Nature of Deception in Negotiation." In *Proceedings of the 1990 Distributed AI Workshop*, October.

Zlotkin, Gilad, and Jeffrey S. Rosenschein. (1991). "Incomplete Information and Deception in Multi-Agent Negotiation." In *IJCAI91*, August, Los Altos, CA: Morgan Kaufmann.