Effective Communication Techniques in 2D Multi-Agent Systems

COMP 3190 Artificial Intelligence

Katrina Dotzlaw

**Abstract**

Communication in multi-agent systems is vitally important for agents to continually update their incomplete views allowing them to efficiently execute actions throughout the system environment. The chosen communication protocol will affect decision making and therefor the level of expected performance. Using a selected set of success criteria which includes the necessity of communication, information space allotment, value calculations and dealing with common communication problems, both cognitive communication techniques and reinforcement learning techniques were compared. After analyzing both cognitive communication techniques and reinforcement learning techniques, and evaluating all techniques using the success criteria outlined, it can be conclusively stated that reinforcement learning is the most effective communication technique.

1. **Introduction**

Communication is a way for the agents in the system to change their own incomplete views by swapping their own information with information belonging to other agents in the system [Xuan and Lesser, 2002]. This then allows agents to mutually discover the global state of the system at a specific time [Xuan and Lesser, 2002].

Multi-agent systems need to communicate information throughout the system. This information can be irrelevant, redundant, or important but unable to be acted upon [Paulos et al., 2019]. The conditions needed by agents to communicate are formulated based on the value of information being communicated [Becker et al., 2009]. The value of information is the expected increase in the value of the best action as a result of receiving the information [Becker et al., 2009].

Typically, the benefits of communication are measured depending on the value of information provided to other agents [Becker et al., 2009]. Optimally, communication policies involve agents choosing the method of communication at each step that maximizes the expected future utility [Becker et al., 2009]. This includes agents choosing not to communicate [Becker et al., 2009]. An agent will not need to communicate if it knows exactly which action it is expected to execute [Xuan and Lesser, 2002].

In this paper, the overall effectiveness of each communication policy will be evaluated based on the following success criteria: the necessity of communication, the space required to store the information, how each policy deals with the common problems with communication and the value of the information transferred. The measurability of each of the four success criteria will first be described. Following this, a discussion of the common communication problems encountered within multi-agent systems and an outline of cognitive approaches and behavioural approaches to communication, respectively will be presented.

1. **Quantitative Criteria Outline**

Points will be assigned to each of the four criteria. If the policy used allows agents to choose when to communicate, the policy receives 1 point. No points are given if it does not. If the space required for the information is low, the policy receives 2 points, if it is medium it receives 1 point. No points are given if the policy requires high amounts of space. If the policy deals with the common communication problems, it will receive 2 points. If it deals with one of the common communication problems, it will receive 1 point. No points are given if it does not deal with any. If the policy does a value calculation, it will be given 1 point. No points will be given if there is no value calculation done.

A total of 6 points can be received. This points system will allow for a definitive answer to the question which communication policy is the most effective, as the answer will be backed up with quantitative data. In the case of a tie, advantages and disadvantages that do not fit into the success criteria will be examined.

1. **Common Problems with Multi-Agent Communication**

The two most common problems associated with multi-agent communication are the Hidden State Problem and the Credit Assignment Problem. Communication problems also arise in situations where agents must interact with their environment [Balch and Arkin, 1994].

* 1. **The Hidden State Problem**

The Hidden State problem occurs when an agent’s next action depends on hidden information i.e the state [McCallum, 1996]. Discovering hidden state requires information about previous actions to determine the state [McCallum, 1996]. However, discovering hidden state requires that agents have a longer training time [McCallum, 1996]. As detailed by McCallum, "A reinforcement learning agent suffers from hidden state if at any time the agent's state representation is missing information needed to determine the next correct action" [McCallum, 1996].

* 1. **The Credit Assignment Problem**

In 1998, Mataric stated that: “The credit assignment problem arises because reinforcement in a distributed system is often provided at a global level and must somehow be divided over multiple agents whose impact differs and varies over time" [Mataric, 1998]. To simplify, credit assignment problems occur when feedback is given at the end of an action sequence [Fu and Anderson, 2008]. It requires the learned to associate the feedback received from earlier actions, and the interdependencies of these actions require the learner to recall their past choices [Fu and Anderson, 2008]. This communication problem involves both an explicit memory encoding process and an implicit reinforcement learning process [Fu and Anderson, 2008]. The explicit memory encoding process requires that memories are rehearsed [Fu and Anderson, 2008]. The implicit reinforcement learning requires the learner to extend credit back to their previous choices [Fu and Anderson, 2008].

In the case of reinforcement learning, the effects of the credit assignment problems differ depending on if implicit or explicit learning is dominant [Fu and Anderson, 2008]. When implicit reinforcement learning was dominant, the selection of the better option in the last choices the learner made was faster that in the first choices they made [Fu and Anderson, 2008]. When explicit reinforcement learning was dominant, the selection of the better option in the first choices the learner made was faster than the last choices they made [Fu and Anderson, 2008].

* 1. **Alleviating Communication Problems**

To alleviate both the credit assignment and hidden state problems, we first consider a suggestion made by Mataric. According to Mataric, the role of communication is to increase the scope of a single agent [Mataric, 1998]. Increasing the scope of a single agent should then make the overall system of agents “temporarily and locally less distributed” [Mataric, 1998], and should help decrease the effects of both the credit assignment and hidden state problems [Mataric, 1998]. With regard to reinforcement learning specifically, Mataric stated that: “In both cases, communication is treated much the same as sensory inputs, thus requiring no additional processing overhead nor a departure from the reinforcement learning framework” [Mataric, 1998]. Hence, the effects of both communication issues would be alleviated. Based on this work done by Mataric, a communication policy that increases the scope of an individual agent would be more effective as it would decrease the negative effects of both the credit assignment and hidden state problems.

It was previously mentioned that an agent’s interaction with the environment can also result in communication issues. In 1994, Balch and Arkin conducted an experiment where they tested the impact of communication on multi-agent robots that participated in tasks where they were required to interact with their environment [Balch and Arkin, 1994]. After conducting their experiment, they concluded two things. Firstly, they concluded that “Communication improves performance significantly in tasks with little environmental communication” [Balch and Arkin, 1994]. Secondly, they concluded that “More complex communication strategies offer little or no benefit over low-level communication” [Balch and Arkin, 1994].

These two conclusions can be applied to the previously mentioned success criteria. If a communication policy has too much environmental interaction, it will be less effective than a policy that does not as the performance improves with a small amount of environmental interaction. The second conclusion noted by Balch and Arkin disqualifies the complexity of a policy from being a part of the effectiveness calculation. If there is no benefit for using a low-level communication policy over a complex communication policy, then there is no impact on the overall effectiveness of the policy.

1. **Cognitive Approaches to Communication**

In most cognitive approaches, agents are faced with multiple types of uncertainty when in a multi-agent system [Carlin and Zilberstein, 2009]. Agents can be uncertain about their own actions, state, and the state of other agents and these other agents’ future actions [Carlin and Zilberstein, 2009]. This uncertainty can be addressed by allowing agents to communicate with other agents in the system [Carlin and Zilberman, 2009].

* 1. **The Myopic Approach**

According to Becker et al., using a myopic algorithm is a prevalent way of dealing with the complexity built-in to finding at least one optimal solution [Becker et al., 2009]. Becker et al. conducted an experiment where they proposed two common myopic assumptions. The first is that “each source of information is evaluated in isolation” [Becker et al., 2009]. This implies that each agent stores its own information that needs to be evaluated. Hence, this policy receives 1 point based on the limited space used. The second is that “a 1-step horizon is used in sequential decision making” [Becker et al. 2009].

The first assumption is also the most basic myopic approach and was used by Becker et al. in their experiment [Becker et al., 2009]. It operates on the assumption that the “value of communication is the difference between the expected value when communicating and he expected value for remaining silent” [Becker et al., 2009]. Since it is the most basic approach, it assumes that other agents never start the process of communicating [Becker et al., 2009]. Another implication of this assumption is that not communicating is a form of communication [Becker et al. 2009]. The second assumption implies that no agent will communicate in the future [Becker et al., 2009].

Since there is a value calculation mentioned however abstract, this policy receives another point. This policy assumes that other agents cannot start the process of communicating, which as previously stated, implies that not communicating is an option. Hence, this policy receives another point. Unfortunately, there is no explicit mention of solving any of the common communication problems and this policy receives no points in that category.

Myopic assumptions could be the best way to estimate the value of information in the case of a single agent, but it is unclear if this is also true in multi-agent settings [Becker et al., 2009]. As per the success criteria outlined previously, the myopic approach mentioned here would receive a score of 3.

* 1. **The Centralized and Decentralized Non-Myopic Approach**

Non-myopic approaches consist of centralized and decentralized policies [Carlin, and Zilberstein, 2009]. Centralized policies identify the decisions of agents regarding the overall system state [Xuan and Lesser, 2002]. In contrast, decentralized policies are required to only assume partial system knowledge in each agent [Xuan and Lesser, 2002]. Each agent must then deal with communication directly [Xuan and Lesser, 2002]. Decentralized policies also correlate with the decisions of already established agents [Xuan and Lesser, 2002]. The difference between these two policies is that the agents using a decentralized policy do not observe the global state automatically, while in centralized policies the global state is used as the starting point [Xuan and Lesser, 2002].

Centralized policies have been proven to be simpler to solve and methodical methods have been developed to obtain those solutions [Xuan and Lesser, 2002]. Decentralized models are advantageous because of their representational power; they represent solutions from multiple points of view [Xuan and Lesser, 2002]. However, once these models acquire decentralized policies, the complexity of the model increases [Xuan and Lesser, 2002]. It must be noted that when decentralized settings have partial observability, resources used for communication can be limited and costly [Carlin and Zilberstein, 2009]. It is also possible that some of the current approaches to decentralized settings with partial observability underestimate the cost of communication which leads to overcommunication [Carlin, and Zilberstein, 2009]. This overcommunication caused by the agents ignore the possibility of other agents postponing or commencing communication [Carlin and Zilberstein, 2009].

It is complicated to compute the value of communication as each agent only receives partial observations and is then required to reason about what other agents are observing and could have to determine the system state post-communication [Carlin and Zilberstein, 2009]. However, this value calculation was determined by Carlin and Zilberstein as an effective method of determining advantages of communicating or the advantages of not communicating [Carlin and Zilberstein, 2009]. Carlin and Zilberstein suggest that it could be more effective to use the value of the centralized policy, as it would be a quicker heuristic [Carlin and Zilberstein, 2009]. Carlin and Zilberstein’s work briefly discusses that while communication can solve problems and execute solutions using a single centralized agent, this both defeats the purpose of having a multi-agent system and assumes that the communication was ubiquitous, which real world communication is not [Carlin and Zilberstein,2009].

The value calculation presented allows the agents to determine when to communicate. Hence, a total of 2 points will be given to both the centralized and decentralized communication policies.

Both the myopic and non-myopic approaches contained a value calculation and allowed agents to determine when to communicate. The myopic approach received an extra point for using a small amount of space. In total, the cognitive approach to communication receives a total of 3 points.

1. **Reinforcement Learning**

Reinforcement Learning is a term describing methods where an agent learns based on a reward-punishment system, where the rewards and punishments are received from the environment [Mataric, 1997]. Instead of receiving reinforcement based on behaviour directly, reinforcement learning allows agents to receive feedback based on the outcome of its actions to a value that images environmental events [Anderson et al. 2002]. Reinforcement learning is typically used to avoid the myopic assumptions [Becker et al. 2009].

Methods of reinforcement learning are typically sorted into two groups. [Hasinoff, 2002]. The first group consists of those that “learn the policy directly, by constructing a value function over states and actions” [Hasinoff, 2002]. The second group consists of those that traverse the policy space directly [Hasinoff, 2002].

* 1. **Learning using the Value over States Function**

When using an observable Markov Decision Process, reinforcement learning techniques from the first group – directly learning the policy -- are the most popular [Hasinoff, 2002]. Gradual learning (or the value over state function) uses a greedy policy to find an optimal solution; the closer the Q-values get to the exact value, the closer the greedy policy is to the optimal solution [Hasinoff, 2002]. Memoryless policies tend to be the simplest approach as they ignore any complications and just apply Q-learning or a similar technique [Hasinoff, 2002]. Hasinoff notes that “even the best memoryless policies can have poor performance, particularly if explicit information gathering is required of the agent” [Hasnioff, 2002]. There are also hierarchical reinforcement learning techniques which have immense potential for “gaining computational leverage in order to solve large-scale decision problems” [Hasinoff, 2002]. Of course, there are problems with this method. One such problem is “managing the transfer of control between sub-policies in the presence of noise” [Hasinoff, 2002].

* 1. **Traversing Policies Directly**

Since the second group of reinforcement learning techniques deal with directly traversing the policy space, they do not suffer with the same problems as the first group [Hasinoff, 2002]. They can be extremely slow due to unsuitable constraints imposed on the policy space; regardless how large the policy space is. [Hasinoff, 2002]. It must also be noted that none of the direct policy methods guarantee a good quality solution [Hasinoff, 2002].

* + 1. **Evolutionary Algorithms**

One technique for directly traversing policies is to use an Evolutionary Algorithm [Hasinoff, 2002]. These algorithms assess the policies directly, try for function optimization and are inspired by biological systems [Hasinoff, 2002]. However, evolutionary algorithms do not have as much of a cohesive body of research as some other the more recognizable reinforcement learning techniques [Hasinoff, 2002]. Generally, evolutionary algorithms are slower and do not deal with states that are seldom visited, despite the fact that these algorithms take less memory [Hasinoff, 2002].

* + 1. **Gradient Ascent Methods**

Gradient ascent methods also fall under the roof of direct policy search techniques [Hasinoff, 2002]. As previously mentioned, the direct policy search techniques “require the value of the policy to be a differentiable function” [Hasinoff, 2002]. Provided that requirements are held, estimates of the gradient can be made and used to represent a form of gradient ascent [Hasinoff, 2002]. Previous applications of this concept were slow and only worked on memoryless policies [Hasinoff, 2002]. It must also be noted that the methods of these applications in regard to gradient approximation had high variance [Hasinoff, 2002]. Newer applications have proved to be a more accurate method for evaluating the gradient and these methods were shown to “converge even in the presence of (reasonable) function approximation” [Hasinoff, 2002].

* 1. **Evaluation of Reinforcement Learning**

Reinforcement learning does not require agents to have a complete model of the domain [Becker et al. 2009]. However, avoiding using the complete domain model then requires the agents to use on-line learning which could lead to poor performance [Becker et al. 2009]. Another issue with reinforcement learning is the trade-off made when using the best-known policy and doing deeper environmental exploration [Hasinoff, 2002].

Reinforcement Learning techniques take longer to converge on an optimal solution [Hasinoff, 2002]. Researchers tend to assume that the agent always starts with no knowledge [Hasinoff, 2002]. Often, there is some form of domain knowledge or expert advice that could be given to the agents in the start state [Hasinoff, 2002].

It is important to note that memory-based techniques are useful for clarifying any hidden state [Hasinoff, 2002]. Recent research has suggested that agents can benefit from multi-level memory [Hasinoff, 2002]. However, a large state space could be potentially challenging [Mataric, 1997].

Credit Assignment problems are harder to deal with in multi-agent reinforcement learning than in individual agent reinforcement learning [Anderson et al. 2002]. This is largely because reinforcement can be postponed and could involve the work of other agents [Anderson et al. 2002]. It is also possible that the system of agents does not have any source of immediate reinforcement, and thus credit is delayed [Mataric, 1997].

Reinforcement learning deals with all the common communication problems; the hidden state and credit assignment problems are handled over any method of reinforcement learning, and any environmental problems can be dealt with using a value-over-states function. Thus, 2 points can be added to the policy total. It can also do a type of value calculation if a gradient ascent method is used, so it receives another point. The space constrains mentioned make it unlikely that only a small amount of information can be transferred, but it is unclear exactly how constrained this space is. As a result, this policy will receive 1 point; the value for medium space required. Receiving reinforcement can teach an agent when it should or should not communicate. Therefore, this policy receives another point. Reinforcement learning has a total of 5 points, which were calculated using both the value-over-state function and the traversing directly policy.

1. **Conclusion**

To be able to assess effective communication techniques in 2D multi-agent systems, the various methods were evaluated on their ability to determine the necessity of communication, information space allotment, value calculations and dealing with common communication problems. Based on these success criteria and the quantitative data gathered, it can be concluded that the reinforcement learning communication policy is the most effective. It ended up receiving a total of 5 points. The cognitive communication technique received 3 points. The reinforcement learning technique seems to be the most diverse and the most expandable, as it is modeled after human behaviour.

1. **References**

[Anderson et al. 2002] Anderson, J., Tanner, B. and Wegner, R., 2002, June. Peer reinforcement in homogeneous and heterogeneous multiagent learning. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2002)* (pp. 13-18).

[Balch and Arkin, 1994] Balch, T. and Arkin, R.C., 1994. Communication in reactive multiagent robotic systems. *Autonomous robots*, *1*(1), pp.27-52.

[Becker et al. 2009] Becker, R., Carlin, A., Lesser, V. and Zilberstein, S., 2009. Analyzing myopic approaches for multi‐agent communication. *Computational Intelligence, 25(1)*, pp.31-50.

[Carlin and Zilberstein, 2009] Carlin, A. and Zilberstein, S., 2009, September. Myopic and non-myopic communication under partial observability. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 2, pp. 331-338). IEEE.

[Fu and Anderson, 2008] Fu, W.T. and Anderson, J.R., 2008. Solving the credit assignment problem: explicit and implicit learning of action sequences with probabilistic outcomes. *Psychological research*, *72*(3), pp.321-330.

[Hasinoff, 2002] Hasinoff, S.W., 2002. Reinforcement learning for problems with hidden state. *University of Toronto, Technical Report*.

[Mataric, 1997] Matarić, M.J., 1997. Reinforcement learning in the multi-robot domain. In *Robot colonies* (pp. 73-83). Springer, Boston, MA.

[Mataric, 1998] Mataric, M.J., 1998. Using communication to reduce locality in distributed multiagent learning. *Journal of experimental & theoretical artificial intelligence*, *10*(3), pp.357-369.

[McCallum, 1996] McCallum, R.A., 1996. Hidden state and reinforcement learning with instance-based state identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *26*(3), pp.464-473.

[Paulos et al. 2019] Paulos, J., Chen, S.W., Shishika, D. and Kumar, V., 2019, May. Decentralization of multiagent policies by learning what to communicate. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 7990-7996). IEEE.

[Xuan and Lesser, 2002] Xuan, P. and Lesser, V., 2002, July. Multi-agent policies: from centralized ones to decentralized ones. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 3* (pp. 1098-1105).