Review

Learning to Communicate through Multi-Agent Reinforcement Learning (MARL): A Systematic Literature Review

Ali Beikmohammadi *

Posted Date: 17 April 2024

doi: 10.20944/preprints202404.0813.v1

Keywords: Reinforcement Learning; POMDP; Learn to Communicate; Systematic Literature Review; Quantitative Methods; Qualitative Methods

Review

Learning to Communicate through Multi-Agent Reinforcement Learning (MARL): A Systematic Literature Review

Ali Beikmohammadi

Department of Systems and Computer Sciences, Stockholm University, Borgarfjordsgatan 8, SE-164 07 Kista, Sweden; beikmohammadi@dsv.su.se

Abstract: Recent years have witnessed significant advances in reinforcement learning (RL), which has registered remarkable success in solving various sequential decision-making problems in machine learning. Most of the successful RL applications, e.g., the games of Go and Poker, robotics, and autonomous driving, involve the participation of more than one single agent, which naturally fall into the realm of multi-agent RL (MARL), a domain with a relatively long history, and has recently re-emerged due to advances in single-agent RL techniques. Although, typically, the communication protocol between agents is manually specified and not altered during training, recently, some papers have shown signs of trying to emerge a communication between agents on the one hand and, on the other hand, to understand what is exchanged between agents. So, there is a growing body of literature on this topic which includes qualitative and quantitative studies and the ones that apply mixed methods. This study presents the scoping review of the methodological strategies undertaken in a total of 16 research articles. The results present the critical appraisal of quantitative methods in terms of validity and reliability and for qualitative methods considering four trustworthiness factors. In the end, relevant insights are further explored with implications and reflections on how they can benefit one's research in the field.

Keywords: reinforcement learning; POMDP; learn to communicate; systematic literature review; quantitative methods; qualitative methods

1. Introduction

1.1. Background

Reinforcement learning (RL) is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. The computer employs trial and error to come up with a solution to the problem. To get the machine to do what the programmer wants, the artificial intelligence gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward. Although the designer sets the reward policy–that is, the rules of the game–it is given the model no hints or suggestions for how to solve the game. It's up to the model to figure out how to perform the task to maximize the reward, starting from totally random trials and finishing with sophisticated tactics and superhuman skills [1].

Recent years have witnessed astonishing advances of RL in many prominent sequential decision-making problems, such as playing the game of Go [2,3], playing real-time strategy games [4,5], robotic control [6,7], playing card games [8,9], and autonomous driving [10], especially accompanied with the development of deep neural networks (DNNs) for function approximation [11].

Intriguingly, most of the successful applications involve the participation of more than one single agent/player, which should be modeled systematically as multi-agent RL (MARL) problems. Specifically, MARL addresses the sequential decision-making problem of multiple autonomous agents that operate in a common environment, each of which aims to optimize its own long-term return by interacting with the environment and other agents [12]. Besides the aforementioned popular ones, learning in multi-agent systems finds potential applications in other subareas, including cyber-physical systems [13,14], finance [15,16], sensor/communication networks [17,18], and social science [19,20].

Largely, MARL algorithms can be placed into three groups, fully cooperative, fully competitive, and a mix of the two, depending on the types of settings they address. In particular, in the cooperative setting, agents collaborate to optimize a common long-term return; while in the competitive setting, the return of agents usually sum up to zero. The mixed setting involves both cooperative and competitive agents, with general sum returns.

In all MARL problems, communication is a fundamental aspect of intelligence, enabling agents to behave as a group, rather than a collection of individuals. It is vital for performing complex tasks in real-world environments where each actor has limited capabilities and/or visibility of the world. In any partially observed environment, the communication between agents is vital to coordinate the behavior of each individual.

While, in RL-based works, the model controlling each agent is typically learned via RL, the specification and format of the communication is usually pre-determined. For example, in robot soccer [21], the bots are designed to communicate at each time step their position and proximity to the ball. It is very clear that if this communication protocol between agents or between agent and human is predetermined, it could not be optimal across all tasks. Hence, today, with the rapid advances in machine learning in recent years, the goal of enabling intelligent agents to communicate with each other and with humans, rather than relying on explicit supervision, is turning from a hot topic of philosophical debates into a practical engineering problem and is often considered a prerequisite for developing a general AI. For an extensive overview of earlier work in this area, we refer the reader to [22,23].

However, there is a missing link in these studies. Specifically, the main drawback of researchers' efforts is the lack of consideration of communication issues such as delay, communication cost, congestion, and packet limit, especially for communication between agents. In this regard, it seems that developing a new realistic situation considering co-optimization of communication from the perspective of telecommunications network issues and performance in partially observable Markov decision process (POMDP) tasks based on RL algorithms is vital.

1.2. Objective

As described in the background, there has been an increasing tendency to focus on learning to communicate through MARL in AI. However, there is a deficiency of the literature review in the research area. Especially researchers tend to neglect the research methodology behind the scene. As it is a methodologically driven systematic literature review, the aim is to provide an overview of the methodological approaches currently undertaken in scientific studies involving learning to communicate with MARL. Particular interest is given to understand how the studies guarantee validity and reliability for quantitative methods and trustworthiness for qualitative methods.

To be more specific, the underlying research questions are:

- What are the research methodologies applied in researches regarding learning to communicate with MARL?
- What kind of problems or challenges are there when applying these research methodology designs?

Roadmap. The remainder of this paper is organized as follows. In Section 2, the methods followed in conducting this systematic review are described, from searching for relevant papers to analyzing strategies. In Section 3, the results of the analysis are reported. And finally, in Section 4, the problems and challenges of the current methodologies when researching learning to communicate with MARL are discussed.

2. Methods

For this overview, the methodological approach is a systematic literature review, which is a stand-alone literature review of research done in a given topic, conducted in a systematic, rigorous manner, without collecting or analyzing any new or original data [24].

In this exploratory study, we examine articles published between 2016 and 2021 (except for one, which was later added via backward and forward searches.) This specific time range has been chosen by taking the "Learning to Communicate with

Deep Multi-Agent Reinforcement Learning" paper [28] as a reference point. To achieve an abundance of published work in the field, papers registered in the academic search engine Google Scholar will be the primary resource. To ensure that the selected articles meet recognized standards of quality, they were searched in the Scopus database as well, as it only archives peer-reviewed papers. We did not search for literature on Scopus instead because Google Scholar has the advantage of being able to search keywords in the text, which was beneficial in the case where the same query applied to Scopus would produce very few results (even in some cases nothing).

Note that it is widespread in the AI/ML field; due to the rapid research progress, papers are made available before review. Sometimes, even while they are under review, many other works refer to them. In this regard, although it has been tried to use only reviewed peer-reviewed papers as much as possible, this is not a definitive criterion for selecting papers, and appropriate papers, which are highly cited works in this field, have been considered in this report.

2.1. Searching for Relevant Papers

The main component of the created queries for the literature search process is ("POMDP" OR "multi agent reinforcement learning" OR "multi-agent reinforcement learning" OR "MARL") AND "learn to communicate", which will further be referenced as mainQRY. These key terms are targeting research that concerns specifically the field of MARL and its community.

Table 1 gives a summary of all the executed queries on November 15, 2021, at each stage. It should be noted that the study period is between 2016 and 2021. In Google Scholar, the execution of the following initial query, which was done before any specific literature search, mainQRY produced 216 results, and in Scopus, mainQRY produced 13 results.

Different types of data could be generated in a research study; however, the two main types are qualitative and quantitative. Thus, the search procedure and the respective analysis of the literature initially follow this main two-fold categorization as well as some related keywords. The search was conducted in two stages (see Table 1).

Table 1. Summary of the searching process for finding papers for the review, published between 2016 and 2021. (mainQRY = "POMDP" OR "multi agent reinforcement learning" OR "multi-agent reinforcement learning" OR "MARL").

Stage	Query	Search Engine/Database	Search Results (# Papers)
Initial stage	mainQRY	Google Scholar	201
	mainQRY	Scopus	13
Final stage	mainQRY AND ("participants" OR "interview" OR "questionnaire" OR "quantitative" OR "qualitative")	Google Scholar	69
	mainQRY AND ("participant*" OR "interview" OR "questionnaire" OR "quanti*" OR "quali*")	Scopus	2

The final stage was performed to have an overall idea of how many targeted papers use quantitative or qualitative or mixed methods in their study. As Google Scholar does not support truncation symbols (the abbreviation "quanti*" would have results that included quantitative, quantitatively, and other), the full words were used, knowing that without quotation marks the algorithm takes into account words that have the root quantitative or qualitative, e.g., quantitatively. However, in this case, the search did not include words that have the root quant or qual. These terms may seem restrictive, as not all articles explicitly state that they followed a qualitative or quantitative approach. However, given that the range of methodologies is extensive and there is no delimited terminology used in the field, these keywords are restrictive enough to provide a wide range of methodologies, further seen in the Results section.

Table 1 gives the search queries for the final stage and the resulted number of papers. For Scopus, the search queries produced only two results, which further reinforced the necessity of using Google Scholar.

2.2. Inclusion Criteria

The following criteria were followed for deciding whether a paper will be included in the review:

- Publications written in English
- The title and abstract give hints that the topic of the research involves both MARL and Learn to Communicate.
- Has an extensive method section (at least half of a page)
- Exhaustively specifies that a quantitative or qualitative research approach has been used
- Uses collected empirical data or primary data or secondary data for data analysis.
- Regarding qualitative papers, the title or abstract contains a hint that the study has a qualitative nature. Considering the scarcity of these types of papers in the analyzed field, the paper will be reviewed if it mentions interview, participant, questionnaire, ….

2.3. Exclusion Criteria

The exclusion reasons for relevance were:

- Being a literature review study.
- The absence of a method section
- The methodological approach did not correspond with the specified data type (qualitative and quantitative)
- The studies only reference the field of MARL but do not contribute themselves.
- Theoretical, conceptual papers without empirical research
- Short papers/ Workshop papers

By applying inclusion and exclusion criteria, out of 69 found papers, 15 remained relevant for the literature review. Out of 15 chosen papers, 10 were categorized as following a quantitative approach, two as following a qualitative approach, and three as mixed methods. In addition to these 15 papers, it was decided to add another qualitative paper [42] to this collection based on backward and forward searches. As the latter three papers have a relatively equal share in qualitative and quantitative methods, they can be examined from the point of view of both qualitative and quantitative categories. However, in this paper, since the quantitative approach being the most used in the field, they only will be included in the qualitative category.

Out of the 16 papers, 9 have been deposited on the arXiv repository. The remaining articles were published in conferences: Advances in Neural Information Processing Systems, IEEE International Conference on Computer Vision, AAAI Conference on Artificial Intelligence, International Conference on Machine Learning, IEEE International Conference on Robot & Human Interactive Communication, International Conference on Persuasive Technology. Thus, it shows that there is a tendency to publish papers in conference proceedings and publish pre-print papers in the arXiv repository.

2.4. Analysis Strategies

Regardless of whether the papers are categorized qualitatively or quantitatively, in order to gain more knowledge of what each selected paper focuses on and to understand the various contexts that all revolve around trying to learn communication through MARL, the goal of the papers will be investigated. In addition, for the papers, according to their quantity or quality nature, several criteria will be measured, which are introduced below.

For the papers following quantitative approaches, the papers will be analyzed with the two main criteria: reliability and validity. Reliability is about a consistent measurement of the phenomena, which can be translated into reproducibility in data science papers [25]. Highly reliable papers should be able to be reproduced when people follow the instruction and resources, such as code and data. Furthermore, these resources should be easily obtained. Validity is more about how accurately the research reaches the conclusion they want. To be more specific, several criteria have been selected that can measure validity and reliability in this topic.

Validity, Validity can be achieved by showing various materials that can persuade readers to trust the result and its achievements. In AI/ML, the main criterion would be the performance as it is the goal of the papers providing a new approach, but we also put the scalability to support the performance derived in a valid way. Various environments having different specifications would also help to validate the efficiency in various situations. These criteria are for validity:

- **Approach:** In MARL, most works follow one of the well-known algorithms in this field. We checked whether the paper used one of those methods or developed a new model.
- **Performance:** In MARL, the performance of the methods is examined by various methods such as the amount of reward collected, the amount of error. Performance review is very important when you want to compare several algorithms. That is why we evaluate whether the suggestion is always better in any experiment, consistently or not.
- **Number of environments:** The number of environments affects the validity as it can show that the paper's suggestion is consistently winning in multiple cases. If the paper only uses one dataset, we cannot say the paper is highly valid as we cannot expect an undiscovered situation.
- **Scalability:** The number of agents plays a very decisive role when reviewing the results in MARL. As the number of agents increases, the environment becomes closer to the real situation and the proposed algorithm has more validity. We investigated whether the scalability issue is examined (by performing experiments with the presence of a different number of agents) or not. For scalability criteria, we used ordinal scores to show how much each paper keeps the criteria. In this scoring, 0 means that there is no sign of scalability, 1 means there are some intuitions in which symptoms of scalability can be found, and 2 implies an assurance of scalability through testing the algorithm on several environments and/or a different number of agents.

Reliability, Without a shadow of a doubt, making experimental instructions and resources as transparent as possible, so that other people can reproduce the result clearly by following the paper, can ensure reliability. In this regard, the openness of the dataset, algorithm implementation, comparison to the baseline model, and clear experimental setup can be measured in AI/ML.

- **Data availability:** Public datasets/environments that are available for any people to use can make the research more reliable as everyone can restore data in the study. However, when the paper changes a part of a dataset/environment, the reliability will also be reduced as we may need a clear explanation of the changes.
- **Code availability:** Available code will help with reliability as we can see that the code produces the same result. We also can see whether the code and the hypothesis are matching or not.
- **Comparison to others (baseline model):** A comparison to baseline models that are the most recent in the field would be helpful to show reliability. On the other hand, if the paper missed one of the state-of-the-art articles to compare, it might lose reliability. We cannot make sure that the suggestion is better than other recent ones.
- **Experimental setup:** The data science area has a standard research procedure that can be a condition for improving the reliability of the study. We checked whether the paper contains detailed information on data types, how they prepared the data, and the explanation of modeling and evaluation. Ordinal scores are used to show how much each paper refers to details when explaining test settings, data collection, implementation, and evaluation of

the model. In this scoring, 0 means that very little explanation is given, 1 means details are given, and 2 implies an assurance that there is no ambiguity left for reproducing the results.

For the papers following qualitative approaches, the research strategy will be critically analyzed along with the four criteria of measuring trustworthiness [26]: credibility, transferability, dependability, and confirmability. Guba proposed these criteria as strategies that qualitative researchers should follow to ensure that their study is reliable, follows a rigorous research process, and has a pertinent qualitative inquiry. Credibility is to make the confidence of the result true and credible enough. A peer debriefing or thorough interviewing process can affect it. Transferability is whether the result can be transferred or generalized to other settings as well. Data saturation and purposeful sampling can be strategies. Confirmability is about the confidence of the results by being confirmed by other researchers. Finally, dependability is how much the research can be repeatable. Detailed explanation, provision of audit trail, and stepwise replication of the data can increase this value. We used these four criteria as they are widely accepted by many qualitative research communities as valid forms to ensure the quality of the research [27]. In addition, we will investigate the research design applied in each study.

3. Results

Quantitative and qualitative papers are evaluated by the criteria discussed in the Methods section separately. Here, for scalability and experimental setup criteria, as it is defined in Section 2.4, we used ordinal scores to show how much each paper keeps the criteria. Although these ordinals scoring for all papers has been determined by the author of this work, we have tried to avoid bias.

3.1. Quantitative Papers

In the first step, the goals of each quantitative paper are identified. The purpose of identifying goals is to understand how these papers can be divided into different categories based on this. Although the objectives of the papers are shown in detail in Table 2, by examining the papers, it can be seen that their goals can be generally divided into two parts. Several articles seek to learn to communicate while learning another POMDP task at the same time as the main task. Specifically, this approach has begun with “Learning to communicate with deep multi-agent reinforcement learning” paper [28]. On the other hand, the rest of the articles are focused on finding a meaningful communication protocol between agents as an emerging language, and in fact, the main task in this approach is only to learn a meaningful communication between agents and there is no other task as the responsibility of agents to learn at the same time [29].

On the other hand, if we want to classify the papers according to the basic methods that exist in RL and machine learning, it can be concluded from Table 2 that although [30,34] have tried to find a new MARL based approach as a solution method, most papers develop their approach based on existing methods, among which articles that use recurrent neural networks such as LSTM, B-LSTM, and GRU make up the majority of studies. The importance of this issue goes back to the fact that validating articles that use known approaches is an easier task for researchers than to validate a new approach that has not yet been fully explored. It is also important to note that the majority of studies, with the exception of three papers, [32,35,36], are focused on only one method, not combining several techniques.

Table 2. Goals and approaches in the selected quantitative studies.

Paper	Goal	Approach
Foerster et al. [28]	Learning a binary (in execution mode) communication protocol	DRQN
Jorge et al. [29]	Learning Guess who? by two agents (asker, answerer)	Based on [28]
Sukhbaatar and Fergus [30]	Learning continuous communication between a dynamically changing set of agents for fully cooperative tasks	New model

Havrylov and Titov [31]	Learning to communicate with sequences of discrete symbols (referential game)	LSTM
Das et al. [32]	Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning	VGG-16, LSTM
Mordatch and Abbeel [33]	Formulate the discovery of the action and communication protocols for their agents jointly as a reinforcement learning problem	LSTM
Jiang et al. [34]	Modeling multi-agent environment as a graph	New model
Celikyilmaz et al. [35]	Addressing the challenges of representing a long document for abstractive summarization by deep communicating agents in an encoder-decoder architecture	B-LSTM, Attention
Das et al. [36]	Proposing an architecture for multi-agent reinforcement learning that allows targeted continuous communication between agents via a sender-receiver soft attention mechanism and multiple rounds of collaborative reasoning.	Actor-critic, GRU
Cogswell et al. [37]	Developing an implicit model of cultural transmission and compositionality in deep neural dialog agents, where language is transmitted from generation to generation because it helps agents achieve their goals	Based on [38]

The most important point is the criterion for evaluating the superiority of a method which can be another way to check the validity of the research. As mentioned in Table 3, the most important criterion in tasks based on RL is normalized rewards. The next criterion as a measure of performance is the rate of win or failure. In some articles, mean errors, losses, and accuracy are also used as other criteria for measuring the algorithm. On the other hand, due to the fact that in RL studies, the choice of criteria is highly dependent on the choice of the experimental environment, sometimes researchers have to define the custom metrics [29,34,35]. It is obvious that the more clearly the metrics can cover all the important aspects, the higher the validity of the research.

Table 3. Summary of performance metrics (data analysis methods) used in the selected quantitative studies.

Performance Metrics	Paper (s)
Normalized rewards	Foerster et al. [28], Das et al. [32], Mordatch and Abbeel [33], Jiang et al. [34]
Failure rates/ Win rates	Sukhbaatar and Fergus [30], Havrylov and Titov [31], Das et al. [36]
Mean error	Sukhbaatar and Fergus [30],
Loss	Havrylov and Titov [31]
Accuracy, precision, and recall	Cogswell et al. [37]
Custom measure	Jorge et al. [29], Jiang et al. [34], Celikyilmaz et al. [35]

Another way to check the validity of research is to experiment with different environments and/or with a different number of agents. In this regard, two factors, the number of tested environments and scalability, have been investigated as the last hints for having a validity criterion using Table 4. While we know for sure that the greater the number of simulated environments on which the proposed method has been investigated, the higher the validity of the research, with more than half of the papers examining only their method on one environment [29,31–35]. They increase the risk of bias in the results. However, Authors in [30] and [36] have performed experiments on five and four environments, respectively, to convince researchers of the validity of the proposed method.

On the other hand, half of the studies have not attempted to demonstrate scalability, which means that their proposed method can be used in a more realistic environment or with the presence of more agents. However, the other papers gave researchers more confidence in scalability by examining the effects of noise during communication, examining the effect of increasing the number of agents, and examining changing the difficulty of the experimental environment. Specifically, [36] has been examined the presence of the ten agents, and on the other hand, has been examined in four different environments, so that it has provided high confidence in the scalability of its proposed method. It is worth mentioning that the existence of an open system with heterogeneous agents can be a representation of the validity of the method, which is not seen in any of the papers.

In the following, we intend to examine the quantitative papers in terms of reliability. In this regard, four factors, the availability of data or simulated environment, the availability of code, comparison of results with basic models or other possible configurations of the proposed model, as well as paying attention to details when explaining the experimental setup have been reviewed in the studies. Data availability can be mentioned as the first factor in checking reliability.

Table 4. Summary of assessing scalability and number of environments in the selected quantitative studies.

Paper	Number of Environments (Ordinal)	Scalability (Ordinal (0-2))
Foerster et al. [28]	2	1
Jorge et al. [29]	1	0
Sukhbaatar and Fergus [30]	5	2
Havrylov and Titov [31]	1	0
Das et al. [32]	1	0
Mordatch and Abbeel [33]	1	0
Jiang et al. [34]	1	2
Celikyilmaz et al. [35]	1	2
Das et al. [36]	4	2
Cogswell et al. [37]	2	0

As can be seen from Table 5, in RL there are two general parts of the data set and the simulation environment. Data sets are used to train neural networks in some methods. On the other hand, the simulation environment means an environment in which the agent tries to learn and improve its performance to perform a specific task by interacting with it. Although half of the papers did not require the use of datasets to pre-train neural networks, the rest of the works attempted to use existing known datasets with higher reliability rather than creating new ones.

In terms of using existing simulation environments or implementing a new simulation environment, the articles have considered both approaches according to their needs. To put this in terms of reliability, using well-known and standard data sets and simulation environments will undoubtedly help future researchers reproduce the results quickly.

Table 5. Summary of data collection methods and environment used in the selected quantitative studies.

Data availability	Paper(s)
Create new dataset	Cogswell et al. [37]
Used public dataset	Foerster et al. [28], Havrylov and Titov [31], Celikyilmaz et al. [35], Das et al. [36]
Implementing a new environment	Foerster et al. [28], Jorge et al. [29], Sukhbaatar and Fergus [30], Das et al. [32], Mordatch and Abbeel [33]
Using the existing environment	Sukhbaatar and Fergus [30], Jiang et al. [34], Das et al. [36], Cogswell et al. [37]

The availability of code is also one of the most important ways to ensure the ability to reproduce results. As shown in Table 6, while some studies, [28,37], allow researchers to reproduce the results as well as to check the compatibility of the code with the hypotheses by inserting pseudocode in the text of the paper as well as by making the code available, half of the papers did not use either of these methods and faced serious challenges for researchers to re-implement their works.

Table 6. Summary of code availability status for the selected quantitative studies.

Code Availability	Paper(s)
Not available	Havrylov and Titov [31], Das et al. [32], Mordatch and Abbeel [33], Celikyilmaz et al. [35], Das et al. [36]
Provided pseudo-code	Foerster et al. [28], Cogswell et al. [37]
Available	Foerster et al. [28], Jorge et al. [29], Sukhbaatar and Fergus [30], Jiang et al. [34], Cogswell et al. [37]

One way to measure the reliability when studying a paper is to compare the proposed method with baseline models as well as the state-of-the-art models. By doing this comparison, the reader can be sure that the proposed method works better than the existing methods. As shown in Table 7, half of the articles, especially [35], compared their method with other methods.

Another way to compare is to compare the performance of the proposed method with other modified versions of your method. The purpose of this work is to ensure the optimality of the proposed method compared to other close methods that have been done in most papers of this type of comparison.

Table 7. Summary of assessing experimental setup and number of comparisons in the selected quantitative studies.

Paper	Number of Comparisons with Baselines Model (Ordinal)	Number of Internal Comparisons (Ordinal)	Experimental Setup (Ordinal (0-2))
Foerster et al. [28]	1	4	2
Jorge et al. [29]	0	5	1
Sukhbaatar and Fergus [30]	3	0	2
Havrylov and Titov [31]	0	4	0
Das et al. [32]	0	2	1
Mordatch and Abbeel [33]	0	2	0
Jiang et al. [34]	3	2	2
Celikyilmaz et al. [35]	7	7	1
Das et al. [36]	4	3	1
Cogswell et al. [37]	5	0	0

The last way refers to details when explaining test settings, data collection, implementation, and evaluation of the model. The more detailed a task is, as in [34], the less ambiguity there is to reproduce the results. We reviewed the studies and found that almost no specific pattern and standard can be found in this case.

At the end of this subsection, it should be noted that none of the papers studied mentions the limitations, which we will discuss in more detail in the Discussion section.

3.2. Qualitative Papers

Table 8 presents a summary of the research strategies identified in the selected qualitative studies. For each paper, the research strategy was defined along with the goal. The most used strategy is quasi-experimental research. A quasi-experimental strategy is defined as an approach to the design of experiments that allows causal inferences to be made despite the absence of procedures for randomly allocating research subjects to experimental conditions. Five papers [39–43] were following the principles of this approach; however, none of the papers explicitly mentioned it.

The following research design is an experimental study defined as a way of assessing causal relationships by, in its simplest form, randomly allocating 'subjects' to two groups and then comparing one (the 'control group') in which no changes are made, with the other (the 'test group') who are subjected to some manipulation or stimulus. Both quasi-experimental studies and experimental ones seek to answer questions of how? and is it possible? usually having an already predetermined hypothesis to validate or to falsify.

Table 8. Research strategies used in the selected qualitative studies.

Research Design	Paper	Goal
Quasi-experimental	Tucker et al. [39]	To investigate human judgments of the robot, agent, or human actions using a dynamic survey
	Strouse et al. [40]	To test how effectively the FCP agents collaborate with humans in a zero-shot setting
	Miura et al. [41]	To investigate whether using legibility as an objective would improve the interpretability of agents' goals by humans
	Woodward and Wood [42]	To evaluate if the proposed POMDP representation produces robust robots to teacher error, (that can accurately infer task details, and that are perceived to be intelligent.)
	Wang et al. [43]	To investigate the impact of a robot's embodiment, its explanation, and its promise to learn from mistakes on trust and team performance
Experimental study	Buehler et al. [44]	To evaluate the benefits of the assistive communication on task performance between robot and human

Credibility is considered as one of the critical criteria of trustworthiness for guaranteeing the internal validity of a research study, in which they seek to ensure that their study measures or tests what is intended. To evaluate the credibility of the studies, an outline of the data collection methods will be needed. Table 9 summarizes the data collection methods that have been identified in the qualitative studies. For each data collection method, papers are enumerated, along with details on the number of participants or the chosen setting. Moreover, the method of recording is also described for each article.

The most used data collection method is an Online survey, as a qualitative research tool [45], applied by five studies [39–43]. This method is followed by an Observational study [44]. Half of the papers [40,43,44] have more than one data collection method using the Questionnaire as a complementary data collection method, which supports triangulation as a provider of credibility. By having multiple data sources, the findings that are defined from the study will be supported. Another form of boosting credibility is the number of participants. The highest number of participants is exhibited in Tucker et al. 's online survey [39]. Having a high number of participants, however, makes the data analysis process much more cumbersome. Another essential detail regarding participants is their recruitment strategy—all of the papers mentioned how they recruited the participants.

Transferability is another criterion for ensuring trustworthiness, which concerns external validity, e.g., the extent to which the results of a study could be replicated in other circumstances. Shenton, however, emphasizes that one is unable to determine if the findings of a study could be applied to other scenarios, as the studies are inherently conducted on a small sample of the population or in a particular situation. One way of making the task easier or more transparent is to acknowledge and describe the boundaries of one's study, e.g., number of participants, data collection methods, data collection sessions, and the time over which it was collected. The number of participants and the data collection methods have been discussed in the previous part and can be seen in Table 9. All papers have specified their data collection methods. Regarding the time period over which the data was collected, none of them mentioned the date. However, even if the date is not mentioned, the time frame between conducting the research and publishing a paper is usually

doi:10.20944/preprints202404.0813.v1

relatively small. Thus, one can have the year when the article was published as a reference. Lastly, the number and the length of each data collection session were mentioned in four papers [39,40,43,44].

The dependability criteria address the issue of reliability, which entails a technique showing that if the work were repeated, in the same context, with the same methods, and with the same participants, similar results would be obtained. In qualitative studies, this is problematic, as one of the inherent characteristics of this kind of study is that the researched phenomena can quickly change. Describing how the study was conducted, what was done to ensure reliability in the findings, and the effectiveness of the methods is one way of addressing the problem.

Regarding the research design and its implementation, it has been discussed earlier. The reflection on the methods used will be discussed in the next section, and finally, the data gathering will be analyzed. Surprisingly, all the papers have used Content Analysis as the data analysis method, defined as "*a term that refers to a variety of methods for analyzing text, usually, in a quantitative way that involves counting, coding, comparing, contrasting, and categorizing the elements (most typically words) forming a corpus of textual data*" [46]. But none of them explicitly mentioned it.

Table 9. Summary of data collection methods used in the selected qualitative studies.

Data Collection Method	Paper	Details	Method of Recording
Online survey	Tucker et al. [39]	253 participants via Amazon Mechanical Turk	Online answers
	Miura et al. [41]	26 participants via Amazon Mechanical Turk. The only requirement for participation was the ability to read English.	Online answers
	Woodward and Wood [42]	26 participants Consisting of undergraduate and graduate students ranging in age from 18 to 31 with a mean age of 22. Four of the participants were randomly selected for the "human robot" role, leaving for the "teacher" role.	Online answers
Online survey + Questionnaire	Wang et al. [43]	61 participants from a higher-education military school in the United States 14 women, 39 men, age range: 18-23	Online answers
	Strouse et al. [40]	114 participants from Prolific, an online participant recruitment platform 37.7% female, 59.6% male, 1.8% nonbinary; median age between 25–34 years. At the end of the study, an open-ended question for feedback on participants' partners.	Online answers
Observational study + Questionnaire	Buehler et al. [44]	14 participants Participants were randomly divided into two groups, one started with an assisted trial, the other started unassisted. The participants had no prior experience with the task	Recorded actions

Confirmability is the last criterion that ensures trustworthiness, which focuses on objectivity, researcher's biases, method limitations, and the audit trail, i.e., a detailed methodological description. Concerning the choice of methods, only one paper [43] mentions why they chose specific data collection and analysis approaches. Regarding the audit trail, all papers give a thorough description of their method, which was embedded in one of the inclusion criteria for selecting a study for analysis.

Table 10 briefly outlines the limitations that were either mentioned in the studies or identified in the literature analysis. One paper briefly mentions some limitations within the methods used [41]. Most limitations concern the overall process of conducting the study.

Table 10. Summary of limitations in the selected qualitative studies.

Paper	Limitations (Mentioned or Identified)
Tucker et al. [39]	Inadequate analysis Lack of verification of human judgments
Strouse et al. [40]	Inadequate analysis
Miura et al. [41]	It is not always possible to significantly improve legibility over policies maximizing underlying rewards. Their initial evaluations are limited to MazeWorld instances using BST belief update
Woodward and Wood [42]	Inadequate analysis Lack of full explanation of how to collect data Lack of full explanation of the test scenario
Wang et al. [43]	Lack of full explanation of how to collect data
Buehler et al. [44]	Inadequate explanation of the questionnaire Lack of full explanation of how to collect data Lack of full explanation of the test scenario

4. Discussion

Following the analysis of the selected qualitative and quantitative studies, a reflection on some interesting insights that have emerged will continue, along with finding relevance between the methods that have been defined with one's possible future research in the field of MARL or other similar fields. Lastly, a broader reflection on the research area, its societal impacts, and how it aligns with different scientific conventions will be discussed as well.

4.1. Insights from Analyzing the Quantitative Studies

By reviewing quantitative papers, we found that in the field of "learn to communicate with MARL", two goals are implicitly and generally pursued. Some learn to communicate while performing another group task, while others seek to understand what is exchanged between agents as a message. However, none of the papers [28–37] explicitly state not only this goal but also the research question. But fortunately, most of them have stated their specific goals so well. For example, in [36] we can understand that the purpose of the paper is to develop a targeted communication architecture for multi-agent reinforcement learning, where agents learn both what messages to send and whom to address them to while performing cooperative tasks in partially observable environments.

Moreover, in [29,31,32,37] in addition to articulating a specific goal, by expressing their main contributions, the authors of the papers have helped to clarify the purpose of the research. To be more specific, mentioning the main contributions could be helpful to find the answer to some goal-oriented questions in the mind of the reader such as; what are the authors researching? what are the authors trying to discover, prove, or create? how do the authors plan to add value to their academic field? etc.

Regardless of which main goal is considered, special attention has been paid to recurrent neural networks in these articles [28,29,31–33,35–37]. This is important because it is much easier for the reader to verify the proposed methods based on pre-known models, and on the other hand, the validity of the results can be more valid.

However, this alone is not enough, and to fully understand the proposed method, the authors need to give a full explanation of their approach. Unfortunately, this has not been the case in all papers, as in [31,33,37], for example, the proposed algorithm has many ambiguities that make it difficult to re-implement the proposed method. On the other hand, the authors in [28] fully describe their proposed method, so as to ensure the possibility of reproducing the results.

In the case of [30] and [34], although they have proposed a new method which, at first glance, since it is not based on known models, leaves us with doubts about its reliability and validity, but they explain their method in detail, in turn, eliminates this doubt and reassure the reader that their proposed method can be replicated. All in all, it seems that researchers need to pay more attention to fully articulating their proposed method, in order to gain both high reliability and validity by reducing ambiguities.

Moreover, the studies have tried to gain validity criteria by using different performance metrics. While most papers use well-known criteria in the field of RL such as normalized reward [28,32–34], win rate [30,31,36], mean error [30], loss [31], and accuracy [37] to show the validity of their proposed method compared to other works or other assumptions, authors in [29,35] define the custom criteria. In addition to using the normalized reward, [34] uses a custom criterion depending on the simulation environment selected for testing to increase validity, so that for the battle environment the criterion of the number of kills and deaths, for the jungle environment the criterion of the number of attacks, for the routing environment the criterion of the delay and throughput are used.

As another example, in [35] the authors evaluate their system using ROUGE-1 (unigram recall), ROUGE-2 (bigram recall), and ROUGE-L (longest common sequence). As a result, researchers seem to pay more attention to the simulation environment to define the performance criterion and are somehow dependent on it. This can be a challenge when we want to confidently use their method in another environment. On the other hand, other criteria may be considered when choosing a method, such as algorithm execution speed and computational complexity, which have not been studied by any researcher.

Other attempts to obtain a validity criterion relate to the use of different simulation environments and performing experiments by considering various numbers of agents. While the best way to reassure other researchers that the results are valid and that there are no biases in the results seems to be many experiments, most papers, except for [30,36], focus on only one or two simulation environments. Another point in this regard is to experiment with a various number of agents, which in this regard there is an interesting gap between the works, which means that [30,34–36] have investigated several numbers of agents (at least 3 different setups) and [29,31–33,37] consider only one case that gives the reader no insight into the scalability of their proposed method.

On the other hand, the researchers, with the availability of data and code, as well as the description of the experimental setup and doing comparisons with the state-of-the-art models, have tried to achieve reliability. Generally, the easiest way of making the research reproducible is to make their dataset and code open. However, by reviewing the papers, we find that there is no specific pattern among researchers in terms of data availability. [30,34,36,37] typically use existing environments that make it easier for the reader to reproduce the results. However, although the authors in [28–30] have implemented a new environment, by making it available, they have been able to defend their reliability.

Without accessing to code, we cannot make sure that the author did some additional works to make their algorithm shown faster or modified the result without the real code. In terms of code availability, like data availability, there is no specific pattern among the papers, which means that the codes of half of the studies are not available [31–33,35,36], while the other half of the papers have exposed their code in the GitHub to avoid any doubt in the re-implementation process [28–30,34,37]. Furthermore, it is worth mentioning that the authors of [28,37], in addition to making their codes available, also help the reader to facilitate the understanding of their works with the help of putting pseudo-code in the text. In future studies, attempts to make research more transparent, such as the release of code, will have to continue to be reproducible.

Moreover, the studies have tried to gain reliability criteria by doing comparisons with the state-of-the-art models and giving the description of the experimental setup. Although the papers do not follow a single standard in terms of the amount of explanation given about the test settings, and how the data were collected (i.e., complete explanation in [28,30,34], relative explanation in [29,32,35,36], poor explanation in [31,33,37]), all papers provide many comparisons that greatly help the reader's confidence.

To sum up, although no restrictions were mentioned in any of the papers, this area is still very new and has serious shortcomings that need to be addressed. The most important problem is the lack of a suitable simulation environment to examine the impact of communication learning as well as the discovery and control of the protocol of messages exchanged between agents. As there are still no appropriate criteria for assessing communication learning, it is expected that this issue will be addressed in the future.

Also, existing algorithms for real-world use are far from conceivable. This is especially true of algorithms that try to discover only one language between these agents without considering the main task. In general, it seems that in the future, researchers are expected to find a way to optimize algorithms so that both the agent seeks to learn the main task and seeks, at first, to learn how to interact and communicate with other agents with real communication issues in mind, and then, to interpret the behavior of agents with each other and with humans by defining an appropriate metric.

4.2. Insights from Analyzing the Qualitative Studies

By reviewing qualitative articles, we found that the context for learning to communicate through MARL generally pursues two distinct goals from the objectives of quantitative articles. One of these goals is to measure the interpretability and legibility of agent decisions from the human point of view, and the second goal is to try to examine and improve meaningful cooperation between humans and robots. Fortunately, both of these goals are more easily articulated than quantitative works, and their separation is as follows: [41] to the first category and [39,40,42–44] to the second category. Specifically, in [41] the goal is to investigate whether using legibility as an objective would improve the interpretability of agents' goals by humans. So, in general, in qualitative studies, the relationship between humans and agents has been studied more and even this issue continues to the point that the issue of multi-agent gives way to single agent.

Regardless of their purpose, in reviewing the articles, it was found that most of them use quasi-experimental research as research design and most of them seek to examine one or more hypotheses [39–43]. Paper [44], on the other hand, follows an experimental study as a research design. Specifically, in [44], researchers consider the following hypotheses: A communicative assistance concept based on the theory of mind improves joint task performance (H1). The decisions of their communication assistant are similar to the wizard decisions (H2). Compared to alternative communication concepts, their ToM-Com assistant supports a human partner more efficiently, leading to fewer interruptions (H3). Then, to do a study, the 14 participants were randomly divided into two groups, one started with assisted communication, the other started unassisted, which is aligned with an experimental study although not explicitly stated. Note that the problem of not explicitly mentioning the type of research design was seen in all papers. However, all of them have used appropriate research design in their studies.

To evaluate the credibility and transferability of the studies, which are considered for guaranteeing the internal validity and the external validity of a research study, respectively, the data collection methods have been reviewed. While [39–43] have used an online survey, the authors in [40,43,44] have used two sources of data collection, which supports triangulation as a provider of credibility. On the other hand, the number of participants plays a key role in boosting credibility and transferability. However, though the details of the participants have been stated in detail, except for [39] and [40] with 253 and 114 participants, respectively, it seems that the number of participants cannot guarantee credibility and transferability. Most of them, however, have used the online answers method of recording data [39–43], which allows the presence of a large number of participants easily and also helps to have transferability. In fact, this method, since it facilitates the data collection process and also does not require the physical presence of participants which leads the presence of many participants from around the world to be possible, seems to make it very easy to conduct qualitative studies, which will make it easier for future researchers.

As explained in the results section, to acknowledge and describe the boundaries of one's study, in addition to the number of participants and data collection methods, data collection sessions, and the time over which it was collected are known as other ways of making the task more transparent. However, although none of the papers mentioned the date on which the data was collected, the number and the length of each data collection session were mentioned in four papers [39,40,43,44]. For instance, in [39], in the first data-gathering experiments, they asked participants to generate three labels for images taken from the CIFAR10 dataset. They randomly selected 50 images from each of 10 classes and assigned those 500 samples to workers. In total, 51 unique workers completed the task, resulting in 1257 valid annotations. They also have mentioned that the average completion time was 42 seconds [39].

However, it seems that researchers should pay more attention to external validity in their studies. Regarding dependability criteria, which address the issue of reliability, though it is not explicitly discussed in any of the papers, since the characteristics of their studies do not change with time on the one hand, and on the other hand the way the experiments are done is well explained, the works are reproducible.

All the papers have used content analysis as the data analysis method, but none of them explicitly mentioned it [39–44]. For example, in [40], ANOVA is used to compare average team deliveries for each agent partner and in [43], the authors conducted a general linear model analysis with repeated measures and Bonferroni corrections. However, in general, it seems that the analyzes have not been done enough and researchers should pay more attention to this point in the future.

As for confirmability, only the paper [43] mentions why they chose specific data collection and analysis approaches. Also, only in [41], authors explicitly mention some limitations within the methods used. However, all papers give a thorough description of their method [39–44]. Overall, the most important factor preventing studies from being confirmed seems to be the limitations of the overall research process, where, for example, [40] suffer from a lack of full explanation of how to collect data and the test scenario. Therefore, it seems that more comprehensive and in-depth research can be of great help to this field to create a broad insight.

One of the things that can be done in the future in this area is to try to combine the goals mentioned in the quantitative articles with the goals mentioned in the qualitative articles. Specifically, conducting a qualitative study when considering several agents that, in addition to performing the main task, are interacting and communicating with each other and with humans, can link these two parts.

4.3. Reflections

By reviewing 16 quantitative and qualitative papers, we can get the feedback from their study that there is a lot of work to be done in this field and we are looking forward to new creative and innovative researchers who can examine real goals in a real environment in the presence of many agents. Personally, I feel that there is sloppiness in terms of the end goal in learning to communicate through multi agent reinforcement learning, and one should continue to do more to clarify issues in this area. The first step seems to be to provide a tidy, all-purpose framework for the field. And the next step is to create a link between communication in terms of telecommunications networks and communication in terms of interaction between users so that the situation can be closer to the real environment.

4.4. Limitations

The scoping review has three limitations. First, there was no comprehensive search in which the majority of databases were included. Second, it is suggested that scoping reviews involve multiple researchers analyzing the data but given that this report was done by one person, it might have induced some bias in the analysis. Third, the results of the review should be interpreted with caution since they are not provided by an expert in research methodology, but as a result of screening from a researcher in the field of machine learning and reinforcement learning. Fourth, by now, the researcher's focus has been on federated learning [47–49] and single-agent reinforcement learning [50–53], which may not encompass the entirety of multi-agent reinforcement learning (MARL). As a result, the author may have limited experience in MARL.

References

1. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
2. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.
3. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of go without human knowledge. nature, 550(7676), 354-359.
4. OpenAI: Openai five. https://blog.openai.com/openai-five/ (2018)
5. Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W.M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., Ewalds, T., Horgan, D., Kroiss, M., Danihelka, I., Agapiou, J., Oh, J., Dalibard, V., Choi, D., Sifre, L., Sulsky, Y., Vezhnevets, S., Molloy, J., Cai, T., Budden, D., Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Pohlen, T., Wu, Y., Yogatama, D., Cohen, J., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Apps, C., Kavukcuoglu, K., Hassabis, D., Silver, D.: AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/ (2019)
6. Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11), 1238-1274.
7. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.
8. Brown, N., Sandholm, T., & Machine, S. (2017, January). Libratus: The Superhuman AI for No-Limit Poker. In IJCAI (pp. 5226-5228).
9. Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. Science, 365(6456), 885-890.
10. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295.
11. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.
12. Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(2), 156-172.
13. Adler, J. L., & Blue, V. J. (2002). A cooperative multi-agent transportation management and route guidance system. Transportation Research Part C: Emerging Technologies, 10(5-6), 433-454.
14. Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. Computer networks, 101, 158-168.
15. Lee, J. W., & Zhang, B. T. (2002). Stock trading system using reinforcement learning with cooperative agents. In Proceedings of the Nineteenth International Conference on Machine Learning (pp. 451-458).
16. Lee, J. W., Park, J., Jangmin, O., Lee, J., & Hong, E. (2007). A multiagent approach to $ q $-learning for daily stock trading. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(6), 864-877.
17. Cortes, J., Martinez, S., Karatas, T., & Bullo, F. (2004). Coverage control for mobile sensing networks. IEEE Transactions on robotics and Automation, 20(2), 243-255.
18. Choi, J., Oh, S., & Horowitz, R. (2009). Distributed learning and cooperative control for multi-agent systems. Automatica, 45(12), 2802-2814.
19. Castelfranchi, C. (2001). The theory of social functions: challenges for computational social science and multi-agent learning. Cognitive Systems Research, 2(1), 5-38.
20. Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv:1702.03037.
21. Stone, P., & Veloso, M. (1998). Towards collaborative and adversarial learning: A case study in robotic soccer. International Journal of Human-Computer Studies, 48(1), 83-104.
22. Kirby, S. (2002). Natural language from artificial life. Artificial life, 8(2), 185-215.
23. Wagner, K., Reggia, J. A., Uriagereka, J., & Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. Adaptive Behavior, 11(1), 37-69.
24. Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. Sprouts. Concordia University, Canada.
25. Freire, J., Fuhr, N., & Rauber, A. (2016). Reproducibility of data-oriented experiments in e-science (dagstuhl seminar 16041). In Dagstuhl Reports (Vol. 6, No. 1). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

doi:10.20944/preprints202404.0813.v1

26. Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. Ectj, 29(2), 75-91.
27. Pandey, S. C., & Patnaik, S. (2014). Establishing reliability and validity in qualitative inquiry: A critical examination. Jharkhand journal of development and management studies, 12(1), 5743-5753.
28. Foerster, J. N., Assael, Y. M., De Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. arXiv preprint arXiv:1605.06676
29. Jorge, E., Kågebäck, M., Johansson, F. D., & Gustavsson, E. (2016). Learning to play guess who? and inventing a grounded language as a consequence. arXiv preprint arXiv:1611.03218.
30. Sukhbaatar, S., & Fergus, R. (2016). Learning multiagent communication with backpropagation. Advances in neural information processing systems, 29, 2244-2252.
31. Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. arXiv preprint arXiv:1705.11192.
32. Das, A., Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In Proceedings of the IEEE international conference on computer vision (pp. 2951-2960).
33. Mordatch, I., & Abbeel, P. (2018, April). Emergence of grounded compositional language in multi-agent populations. In Thirty-second AAAI conference on artificial intelligence.
34. Jiang, J., Dun, C., Huang, T., & Lu, Z. (2018). Graph convolutional reinforcement learning. arXiv preprint arXiv:1810.09202.
35. Celikyilmaz, A., Bosselut, A., He, X., & Choi, Y. (2018). Deep communicating agents for abstractive summarization. arXiv preprint arXiv:1803.10357.
36. Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019, May). Tarmac: Targeted multi-agent communication. In International Conference on Machine Learning (pp. 1538-1546). PMLR.
37. Cogswell, M., Lu, J., Lee, S., Parikh, D., & Batra, D. (2019). Emergence of compositional language with deep generational transmission. arXiv preprint arXiv:1904.09067.
38. Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017). Natural language does not emerge'naturally'in multi-agent dialog. arXiv preprint arXiv:1706.08502.
39. Tucker, M., Li, H., Agrawal, S., Hughes, D., Sycara, K., Lewis, M., & Shah, J. A. (2021). Emergent Discrete Communication in Semantic Spaces. Advances in Neural Information Processing Systems, 34.
40. Strouse, D. J., McKee, K. R., Botvinick, M., Hughes, E., & Everett, R. (2021). Collaborating with Humans without Human Data. arXiv preprint arXiv:2110.08176.
41. Miura, S., Cohen, A. L., & Zilberstein, S. (2021, August). Maximizing legibility in stochastic environments. In 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) (pp. 1053-1059). IEEE.
42. Woodward, M. P., & Wood, R. J. (2012). Framing Human-Robot Task Communication as a POMDP. arXiv preprint arXiv:1204.0280.
43. Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018, April). Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In International Conference on Persuasive Technology (pp. 56-69). Springer, Cham.
44. Buehler, M. C., Adamy, J., & Weisswange, T. H. (2021). Theory of Mind Based Assistive Communication in Complex Human Robot Cooperation. arXiv preprint arXiv:2109.01355.
45. Braun, V., Clarke, V., Boulton, E., Davey, L., & McEvoy, C. (2020). The online survey as a qualitative research tool. International Journal of Social Research Methodology, 1-14.
46. SAGE: Methods map: Content analysis (2022), http://methods.sagepub.com/methodsmap/ content-analysis, last accessed 8 January 2022.
47. Beikmohammadi, A., Khirirat, S., & Magnússon, S. (2024). On the Convergence of Federated Learning Algorithms without Data Similarity. arXiv preprint arXiv:2403.02347.
48. Beikmohammadi, A., Khirirat, S., & Magnússon, S. (2024). Distributed Momentum Methods Under Biased Gradient Estimations. arXiv preprint arXiv:2403.00853.
49. Beikmohammadi, A., Khirirat, S., & Magnússon, S. (2024). Compressed Federated Reinforcement Learning with a Generative Model. DOI: 10.13140/RG.2.2.32811.45606.
50. Beikmohammadi, A., & Magnússon, S. (2024). Accelerating actor-critic-based algorithms via pseudo-labels derived from prior knowledge. Information Sciences, 661, 120182.
51. Beikmohammadi, A., & Magnússon, S. (2023, May). Comparing NARS and Reinforcement Learning: An Analysis of ONA and Q-Learning Algorithms. In International Conference on Artificial General Intelligence (pp. 21-31). Cham: Springer Nature Switzerland.
52. Beikmohammadi, A., & Magnússon, S. (2023). Human-inspired framework to accelerate reinforcement learning. arXiv preprint arXiv:2303.08115.
53. Beikmohammadi, A., & Magnússon, S. (2023, May). TA-Explore: Teacher-assisted exploration for facilitating fast reinforcement learning. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (pp. 2412-2414).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.