Evaluation of Multi-Agent Systems: Observations from Interactive Games

Rejane Frozza[1, 2♣], Yves Demazeau[1♦], Philippe Gaudé[3], Luis Otávio Alvares[2♥]

[1] Laboratoire LEIBNIZ-IMAG – 46, Avenue Felix Viallet
38031 Grenoble, France – Yves.Demazeau@imag.fr

[2] Instituto de Informática, UFRGS – Av. Bento Gonçalves, 9500
91501-970 Porto Alegre, Brazil – {frozza,alvares}@inf.ufrgs.br

[3] GALILEA Multimédia, 41 bis, Rue Abbé Grégoire
38000 Grenoble, France – phil@galilea.com

Abstract. Multi-agent systems can be used for a variety of applications in several domains. Interactive games are one of these applications. Multi-agent systems applied to games are a rich source of research on interaction, communication, adaptability and management of resources. This paper presents a proposal for criteria and measurements to evaluate multi-agent systems, using a game system as application. These measurements may help to analyse the operation and behaviour of multi-agent systems, helping also to the further design and development of such systems. We believe that the proposed criteria consider important features in a multi-agent environment and, therefore, can be applied to different problems represented as multi-agent systems. An example of the use of these criteria in a game prototype is described, as well as the results achieved.

1 INTRODUCTION

This paper discusses the evaluation of multi-agent systems using interactive games as application. Interactive games are today a source of computational and human research. It is *computational* because of the technical aspects regarding system operation and development and it is *human* because of aspects linked to the interaction between users and the system.

There are only few works about evaluation. The RoboCup research initiative [1] provides a rich environment for analysis and comparison of multi-agent systems. Some works related to this issue can be found in [2], that compares teams and attempts to analyse and generalise the lessons learned; in [3], that analyses, evaluates and understands team behaviour; and in [4], that presents ISAAC, an automated assistant to perform post-hoc analysis of RoboCup teams.

More studies and proposals to compare and evaluate multi-agent systems are required. The work developed on this issue is still in an early stage, the research community needs to be able to compare its results and evaluation can be helpful from the design phase to the usage phase of a system.

♣ CAPES fellow and Student at UFRGS (Brazil).
♦ Research Fellow at CNRS (France).
♥ Professor at UFRGS (Brazil).

This research presents a proposal for criteria and measurements to evaluate game systems developed as multi-agent systems. Criteria can be either general (in relation to the problem) or related to a domain [5]. Here, the domain is represented by interactive games and the problem is either resolution or simulation. These measurements may help to analyse the operation and behaviour of multi-agent systems, may also help to further design and development of such systems. Also, they illustrate the applicability of general multi-agent systems evaluation criteria to games that use multi-agent approach.

This paper is organised as follows: Section 2 presents an overview of the use of benchmarks in multi-agent systems and the features of interactive games. Section 3 presents the evaluation items proposed for interactive games through multi-agent systems. Section 4 presents the application of the several evaluation items in a prototype. Section 5 presents the conclusion and perspectives.

2. MULTI-AGENT SYSTEMS, BENCHMARKS AND INTERACTIVE GAMES

2.1 Multi-Agent Systems

A multi-agent systems [6] consist in a group of agents with interaction ability embedded in the same environment with the aim of performing their tasks. These systems have features related to agent actions, such as cooperation, coordination, communication and organisation, among others.

According to [5], multi-agent systems are made up of models of *agents*, *environment*, *interactions* and *organisations* (VOYELLES approach). These models should be selected in the analysis and design phases and should be adapted to the problem to be solved or to the system to be simulated. The reported work contributes to this method.

Multi-agent systems represent an interesting perspective for the design and development of more powerful and sophisticated environments and tools in order to benefit the increase in complexity of real systems. Thus, according to [7] and [8], they make possible:

- to increase the robustness of a system, preventing failures,
- to increase problem solving execution speed,
- to integrate problem solving methods,
- to integrate knowledge sources.

2.2 Benchmarks

A source for research that has not been much exploited yet, even though it is an interesting one in the multi-agent systems domain, is the issue related to *benchmarks*. The use of benchmarks may help in the design of multi-agent systems because they allow an analysis of a system's behaviour.

In the work by [9] benchmarks are defined as points of reference that allow the comparison of results from different solutions to the same problem and that show the behaviour of a system. For example, in a computer network, benchmarks can be used to measure message transmission time and transmission sequences in the Internet. They can also be used to compare different computer processors, databases and multi-agent systems in order to select the best one.

In the context of this paper, benchmark is considered as a measure to evaluate a system's behaviour. In order to evaluate a system, the relationship (criterion $\rightarrow$ measure) has to be determined and established for each evaluation proposed item. There is no consensus among researchers about a single group of criteria to evaluate multi-agent systems. Research is still in its early stages and present criteria that are occasionally specific as, for example, criteria used

to evaluate only agent architecture, simulation environments or development tools. Some criteria related to multi-agent simulation environments can be found in the work by [10], as criteria related to validation and comparison of multi-agent architecture can be found in the work by [11].

We now propose a new set of criteria as complementary criteria, in addition to those already proposed in other works.

2.3 Interactive Games (Application Domain)

Games can be considered as a support for the establishment of new forms of communication and interaction between computer systems and users. Additionally, they are appreciated by enterprises because they have educational and entertainment qualities that appeal to a large number of people. For example, enterprises could use games to recruit new employees [12]. Game systems enable us to make different types of communication and interactions with users (players), using graphics interface, animated elements, conversation among characters and so on. Therefore, games have important aspects that should be investigated.

Table 1 shows some types of interactive games and their features in relation to multi-agent systems that are considered in this work.

Table 1. Types of Interactive Games

Types of Games	Features
Strategy Games (RTS - Real Time Strategy)	• Mix of tactics, strategies and resource management; • Simulation of military-economic battles in a historical or science-fiction context; • Game characters can be an individual, a small group of individuals or a large number of individuals.
Local (individual sports) and Global (group/team sports) simulation games	• Allow to simulate a certain reality, a certain life; • Sports simulation (e.g. soccer, car races and others); • Sports can be either individual or team sports.
"Sim xxx"-type simulation games	• Represents the simulation of the activity of a large number of people.
Role Games	• The player participates in the story; • The key element is always the individual; • The world progress assumes agents with learning abilities.

The game features that we are interested in, at the level of multi-agent systems, are:

- the number of agents,
- the heterogeneity of agents,
- the scope of goals (global or local),
- the quantification or emergence of a goal.

Characteristics can have emotions, attitudes, mental states, cooperation, interaction and communication abilities and can be conversational. Behaviour represent groups of actions (movements or elementary actions) of game agents as, for example, advance, pull, support, attack and others.

3 EVALUATION CRITERIA

Among video game types, this paper focuses on interactive games of the RTS type (*Real Time Strategy*), Simulation (global and local) and Role Games. These game types are of particular interest in terms of integration of Multi-Agent Systems.

This section presents a first proposal of measures shown – shown on Table 2 – for interactive games with the use of multi-agent systems, from a developer's point of view. The interest here is on the evaluation at the computer level, i.e., at the level of the system itself based on multi-agent systems.

Table 2. Measure Level

Level	In Interactive Games
Criteria at the system level (for developers)	• Impact of characters' contribution on game behaviour • Characters' knowledge level • Success of character's activity • Level of interaction among characters • Deliberation degree

For each proposed criterion, a *weight (W)* can be used, representing the importance of this criterion in relation to global analysis. In this paper, we use $0 \leq W \leq 1$.

The next formula is applied as *global evaluation formula*:

$$E = \sum_{i=1}^{n} Wi.Vi$$

Where:

- i = number of criteria
- Wi = weight of the criterion i
- Vi = evaluation value of the criterion i
- E = evaluation final value

Next is a description of the relationship (criterion → measure) for each proposed evaluation item.

3.1 Evaluation Criteria in Interactive Games

Evaluation items proposed at the system level for tests application are:

Criterion (a): *Impact of characters' contribution (P) on game behaviour*.

This criterion refers to different roles that the character may have during the game. It depends on tasks/activities performed by the characters during the game. This test is significant in RTS (Strategy) games, team simulation (global) games and role games.

How to measure it:

- determine if the features of the characters evolve.
- determine if characters present new behaviour during the game.
- list physical features of characters of the specific game. For example: energy, speed, and liveliness, among others.
- evaluate by time period (t).

For example, in the time t1, P1 perfo rms the task with features energy = 20 and speed = 30, and, in the time t2, P1 finished the task with features energy = 25 and speed = 35. The impact will be the difference among values of the features in relation to a task performed in time period x (in this case, t2-t1). It is necessary to compare several tasks previously performed.

The following rating can be used: Good (3): large evolution, Medium (2): medium evolution, Bad (1): small evolution.

Criterion (b) : *Characters' knowledge level*

This criterion refers to the level of characters' skills and can influence characters' abilities, behaviour and roles. This test is significant (measurable) in RTS type games and role games.

How to measure it :

- determine what the character knows about the other characters and the environment (features abilities).
- determine if character's knowledge is the one the character needs in order to perform its task.

The following rating can be used: Complete (3), Medium (2), Incomplete (1).

Criterion (c) : *Success of character's activity*

This criterion refers to efficiency in characters' plans and skills and can influence the character's evolution behaviour. This test is significant in all games (more in strategy type games – RTS, Global and Local Simulation and role games).

How to measure it : by situations. For example, the character achieves the situation where it prevents another character from entering a room. It depends on achievement of local and global goals.

- determine if actions taken by the character correspond to its intentions (if the character did what it wanted to).

The following rating can be used: Usual (3), Sometimes (2), Difficult (1).

Criterion (d) : *Level of interactions among characters*

This criterion can be applied among pairs of characters and between characters and user.

It depends on:

- type of interaction used: direct (contract net), indirect (blackboard), protocol language.
- interaction goal: conversation alone (without any specific objective); to solve a conflict (dispute a resource) ; information exchange (in order to collaborate, cooperate).

This test is significant in RTS games, global simulation games, "Sim xxx" and role games.

How to measure it :

- determine the time used for each interaction.
- determine interaction goal.
- determine if the interaction takes place occasionally or at a regular basis in the system.

The interaction goal and amount of time used can be analysed for each agent. For example, agent A used x time "only for conversation", y time to "solve a conflict" and z time to "exchange information." Agent B used x1 time "only for conversation", y1 time to "solve a conflict' and z1 time to "exchange information."

The following rating can be used: High (3), Medium (2), Low (1).

Criterion (e): *Deliberation degree*

This criterion is linked to problem solving. It refers to whether agents have decision-making ability and whether agents have mental states. This test is significant in games of the type local and global simulation and role games.

How to measure it : It is possible to determine the amount of use of agents' mental states and to determine whether agents use mental states to make decisions and how often they do it by monitoring their behaviour.

The following rating can be used: High (3), Medium (2), Low (1).

The user can change the weight of each criterion for his/her application. But, we suggest the values presented at Table 3. These values represent the importance of each criterion for each type of game. In our case, our values have been naturally been discussed through interactions with game developers and users, and we do not see by now any other way to help to define them.

Table 3. Suggestion of weight values

	RTS	Global Simulation	Local Simulation	Sim "xxx"	Role
Weight - criterion (a)	0,5	0,8	0,2	0,2	0,8
Weight - criterion (b)	0,5	0,2	0,2	0,2	0,6
Weight - criterion (c)	0,8	1,0	1,0	0,2	1,0
Weight - criterion (d)	0,7	1,0	0,2	0,8	1,0
Weight - criterion (e)	0,2	0,5	0,5	0,2	0,5

3.2 Usefulness of Evaluation Criteria proposed for Game Systems

Each criterion has its usefulness defined in relation to two views:

- the user's (player's) view;
- the developer's view.

Table 4 shows usefulness evaluation criteria under the developer's point of view.

Table 4. Evaluation Criteria Usefulness

Criterion	Usefulness (developer's view)
Impact of characters' contribution on game behaviour	in the conception of characters (assignment of roles to game characters), to try to improve the behaviour of characters in the game through their features.
Characters' knowledge level	for the assignment of abilities, behaviour and roles to characters; if necessary add other characters to the game, the skills of the other game characters have to be known; to assign tasks to new characters.
Success character's activity	the developer can check if the actions taken by the character correspond to its intentions; the developer can check the efficiency of plans and the skills of the characters.
Level of interaction among characters	to control the amount and the cost of interaction; to assign more abilities to characters if the cost of exchanging information is high; to check the behaviour of agents in the system in relation to the interaction and, thus, decide if there is a need to change agents' abilities.
Deliberation degree	to help in the conception of agents and of their mental states, to check if mental states help agents in decision-making.

Table 5 shows usefulness evaluation criteria from the user's point of view.

Table 5. Evaluation Criteria Usefulness

Criterion	Usefulness (user's view)
Impact of characters' contribution on game behaviour	the user can to know the evolution of characters and take advantage of this knowledge for interaction.
Characters' knowledge level	the user can take advantage of the knowledge of the characters to improve interaction during the game.
Success character's activity	the user can observe when the character achieves a goal.
Level of interaction among characters	additionally, it is possible to measure communication between the character and the user.
Deliberation degree	the user can observe the behaviour of the character in the decision-making process.

4 TESTS

In order to validate the criteria proposed in this paper, a game system was used for test application.

4.1 Prototype Description

The prototype is a scene from a role game. It is a dialogue in the supermarket between two agents (a salesclerk and a customer.) The salesclerk tries to sell bad fruits to the customer, motivating the customer to use face and body expressions.

Virtual characters have mental states like emotions and affections. Their emotions are expressed through their face and body expressions. Their affection is characterised by the feelings an agent has in relation to another agent.

The characters' emotions are the following: joy, anger, fear, sadness, surprise and disappointment.

The characters' affections are the following: positive (general impression of the agents), happy for (happiness of an agent when the other is also happy), resentment, like (friendship or animosity), love (or hate), trust (or mistrust), gratitude, fear and admiration.

Affections can influence the agent's current emotional state. For example, if an agent is sad, but meets another agent he loves deeply and who he has not seen for many years, he will become less sad.

The software used to develop the prototype were:

- Python 2.1.1 for Windows, for everything related to the emotions management domain;
- Microsoft Visual C++ 6.0, for everything concerning the 3D domain and file management.

4.2 Experimentation Context of Tests

For collecting information of the dialogue among the agents, we have a file, called Journal. In this Journal, there is information as:

- the messages emitted by agents,
- the messages emitted for agents,
- the change of an agent's emotional state,
- the change of an agent's affection,
- the dates of all events.

An example of a trace and interpretation of a dialogue is presented as follows.

1. Wed Aug 21 16:08:14 2002: [Output Salesman] : *Good Morning.*

Initialisation of the dialogue; the salesclerk agent emits a message; the salesclerk is happy to find a customer.

2. Wed Aug 21 16:08:20 2002: [Output Customer] : *Good Morning.*

The customer agent emits a message; the customer remains impassive. He only looks around.

3. Wed Aug 21 16:08:25 2002:[Output Salesman]: C*an I help you?*

The salesclerk agent emits a message; there is no emotional modification.

4. Wed Aug 21 16:08:30 2002: [Output Customer]: *It is possible.*

The customer agent emits a message; there is no emotional modification.

5. Wed Aug 21 16:08:33 2002: [GEMO Salesman]

Indicates the modification of the emotions of the salesclerk agent.

6. Wed Aug 21 16:08:33 2002: [GAFF Salesman]

Indicates the modification of the affections of the salesclerk agent.

7. Wed Aug 21 16:08:35 2002: [Output Salesman]: *Do you want carpets, fabrics, fruits?*

The salesclerk agent emits a message; the salesclerk has found a potential buyer, he is happy.

8. Wed Aug 21 16:08:38 2002: [GEMO Customer]

Indicates the modification of the emotions of the customer agent.

9. Wed Aug 21 16:08:38 2002: [GAFF Customer]

Indicates the modification of the affections of the customer agent.

10. Wed Aug 21 16:08:38 2002: [Output Customer]: *Do you sell fruits?*

The customer agent emits a message; the customer is surprised, because the salesclerk has what he wants.

11. Wed Aug 21 16:08:44 2002: [GEMO Salesman]

Indicates the modification of the emotions of the salesclerk agent.

12. Wed Aug 21 16:08:44 2002: [GAFF Salesman]

Indicates the modification of the affections of the salesclerk agent.

13. Wed Aug 21 16:08:46 2002: [Output Salesman]: *Of course.*

The salesclerk agent emits a message; the salesclerk is very happy.

This dialogue continues during two minutes. At the end, the salesman is very happy because he sold something to the customer.

Through the interaction between agents presented in this trace, we can say they use mental states all the time and that the mental states expressed by one agent can change the behaviour (mental states) of another agent.

For example, if the customer is angry, the salesclerk can change the conversation or try to calm the customer. This situation represents the ability agents have to make decisions.

4.3 Results of Tests Application

The evaluation allowed us to determine the effective contribution of the evaluation items applied to the prototype.

Next we present each evaluation item and their respective result.

(a) *Impact of characters' contribution on game behaviour:* the agents of the scene do not have physical features, but the conversation evolves during the game. The task of the agents is the conversation. For this item, we assume: value = 1 and weight(criterion) = 0,8.

(b) *Characters' knowledge level*: the agent has a global perception of other agents and of his environment. The perceptions an agent has of other agents are the affections and emotions. We may say, too, that the behaviour of an agent is the result of the observation (perception) of the environment (the rotten fruits, for example) and of other agents (emotion or affection expressed by the other agents). In this case, agents have a level of knowledge that could change the behaviour of the system or of the other agents. Knowledge is represented by the global perception of the agent. For this item, we assume: value = 2 and weight(criterion) = 0,6.

(c) *Success of character's activity:* this test was used on two dialogues: one dialogue between the agents in a successful sale and one dialogue between agents in an unsuccessful sale. For this item, we assume: value = 2 and weight(criterion) = 1.

(d) *Level of interaction between characters*: the purpose of the interaction between the agents of the scene is the *conversation.* The type of interaction used for the conversation is through messages. Interaction situations take place all the time between the agents in the scene. Therefore, the level of interaction between the agents is *high*, but takes place always with the same purpose. For this item, we assume: value = 3 and weight(criterion) = 1.

(e) *Deliberation Degree*: agents have the ability to make decisions and they use their mental states. However, in the scene of this game, agents do not have any choice other than the dialogue. Therefore, agents use their mental states *intensely* during the conversation. For this item, we assume: value = 3 and weight(criterion) = 0,5.

For this example, the global evaluation formula is:

E = weight(a).value(a) + weight(b).value(b) + weight(c).value(c) + weight(d).value(d) + weight(e).value(e)

E = 1.0,8 + 2.0,6 + 2.1 + 3.1 + 3.0,5 = 8,5

The E value represents a rate of evaluation, what can be used for comparison with values generated in other applications that use the same criteria set. This rate of evaluation refers to global behaviour of a multi-agent system.

5. CONCLUSION

This paper has proposed a group of criteria and the possibility to measure each of them in interactive games systems where characters are represented by agents and where even users can be considered agents, every element of the game being part of a multi-agent environment.

Among the difficulties found to determine criteria were the following two:

- even within the single domain of interactive games, criteria and measures proposed do not apply to all different types of games. In order to evaluate each type of game, we have to determine different criteria, even though some criteria can be used in more than one type of game.
- the position of the user in relation to the games systems can influence the proposed evaluation items. The user can be part of the games system or not. Thus, tests might change.

At the level of the designer and developer of a multi-agent system, we can mention the following contribution of the proposed criteria:

- the possibility of evaluating agents' behaviour, abilities and roles in a developed application and, later, improve the conception of new agents;

- they make possible to evaluate the global behaviour of the system and, later, improve the definition of system features, such as interaction and agents' goals;
- the possibility of evaluating the relationship between decision-making and the mental states of agents and decide whether results are advantageous to the system;
- the possibility of checking the efficiency of plans and skills of the agents (if the actions taken by the agents correspond to their intentions).

There are two types of evaluation criteria: intrinsic evaluation criteria (at the system level) and extrinsic evaluation criteria (at the user level). The criteria proposed in this paper are of the intrinsic nature and thus could be applied to general multi-agent systems because they consider important features in a multi-agent environment. Some types of tests at the user level that can be studied in game systems are the following: test of body expressions, test of face expressions and test of character's roles. However, these tests and their utility and application were not discussed in this paper.

We intend to apply the evaluation criteria proposed in this paper in other game systems, because there is an increasing need for analysis and evaluation of systems. Thus, we think that soccer game can be also used as a research platform for applying evaluation criteria of multi-agent systems.

REFERENCES

[1] H. Kitano; M. Asada; Y. Kuniyoshi; I. Noda; E. Osawa. RoboCup: The Robot World Cup Initiative. In: International Conference on Autonomous Agents, 1., 1997, Marina del Ray, Califórnia. Proceedings... [S.l.:s.n.], 1997.

[2] S. Marsella; J. Adibi; Y. Alonaizan; G. Kaminka; I. Muslea; M. Tambe. Experience acquired in the design of robocup teams: A comparison of two fielded teams. Journal of Autonomous Agents and Multi-Agent Systems, special issue on "Best of Agents'99", vol 4, p. 115-129. 2000.

[3] M. Tambe; T. Raines; S. Marsella. Agent Assistants for team Analysis. AI Magazine. 2000.

[4] T. Raines; M. Tambe; S. Marsella. Towards automated team analysis: a machine learning approach. Third International RoboCup Competitions and Workshops. 1999.

[5] Y. Demazeau. VOYELLES. Rapport d'Habilitation à Diriger des Recherches. Grenoble:INPG, Avril 2001.

[6] M. N. Huhns; L. M. Stephens. Multi-Agent Systems and Societies of Agents. Lect ures on multi-agent systems. Chapter 2. MIT Press. 1999. p. 28-79.

[7] A. H. Bond; L. Gasser. An Analysis of Problems and Research in DAI. In: BOND, Alan H.; GASSER, Les (Eds.). Readings in Distributed Artificial Intelligence. San Mateo: Morgan Kaufman, 1988.

[8] N. R. Jennings; J. A. Pople. Design and Implementation of ARCHON's Coordination Module. In: Workshop on Cooperating Knowledge-Based Systems. Proceedings... Keele:[s.n.], 1993. p. 61-82.

[9] M. Grosse. Real-Time Coordination in Multi-Agent Systems towards Benchmarks. Odense:University of Southern Denmark, 2000. Master thesis.

[10] M. Reddy; J. Moon. Development and evaluation of multi-agent and agent-based simulation environments. In: International Workshop on Decentralized Intelligent and Multi-Agent Systems (DIMAS95), 1.. Proceedings... Krakow (Poland): University of Mining and Metallurgy, 1995. p. 393-401.

[11] S. Cavalieri; M. Taisch; M. Garetti; M. Macchi. An Experimental Benchmarking of two Multiagent Architectures for Production Scheduling and Control. In: International Workshop on Intelligent Manufacturing Systems, 1., Proceedings... Lausane, 1998. p. 133-154.

[12] L. Bernaert. Le recrutement en ligne entre dans le jeu. Le Monde Interactif (Journal). Mercredi 13 juin 2001.