Improving Players' Assessment in Crisis Management Serious Games: The SIMFOR Project

Ali Oulhaci, Erwan Tranvouez(✉), Sébastien Fournier, and Bernard Espinasse

Aix Marseille Université, CNRS, LSIS UMR 7296, 13397, Marseille, France
{ali.oulhaci,erwan.tranvouez,sebastien.fournier,
bernard.espinasse}@lsis.org

Abstract. Serious Games (SG) are more and more used for training in various domains, but notably in crisis management. In order to improve training results, learner assessment can provide insights on what went right or wrong during a training session. Such assessment is more complex when actors' individual actions must be considered, but also the results of their interactions (collective actions). Such interactions can either be engaged with real or simulated players, through adaptive dialogues immersing players in the different ways (actions, procedures, …) to manage a crisis. This paper presents a multi-agent simulation and assessment approach of SG players, targeting the management of distributed and heterogeneous information (in nature or source) based on the concept of Evaluation Space allowing the production of individual and collective assessments. This approach is developed and illustrated on the SIMFOR SG dedicated to crisis management.

Keywords: Serious game · Learner assessment · Multi-agent system · Agent based simulation · Crisis management

1 Introduction

The growing interest for Serious Games (SG), especially for training, has raised new needs in terms of learners' assessment [1] and behaviours simulation [2]. SG aims at immersing learners as players in a simulated environment improving thus their motivation and involvement by having players learning by doing [3]. SG can use simulation to reproduce a complex or expensive phenomena (physics, natural disaster, etc.) or when there is a high number of actors, to simulate the human actors' behaviors (called Non-Player Characters or NPC).

The goal of the SG is to teach one or more skills to one or more actors (players). Each SG answers in general to a particular training goal, in relation to a specific training context related to trades (crisis and risk management [3–5], firemen operations [6], …), or a specific body of knowledge (school or university courses), or even of social skills (conflict management, cooperation, etc.) [7]. The qualitative or quantitative measure of the success or failure of learning may at the same time implement *ad hoc* or generic

N. Bellamine Ben Saoud et al. (Eds.): ISCRAM-med 2015, LNBIP 233, pp. 85–99, 2015.
DOI: 10.1007/978-3-319-24399-3_8

solutions. This paper presents how automated assessment can improve the learning objective of a multi-player (and multi-skills) SG, as applied to Crisis Management based on a Multi-Agent Systems (MAS).

We first discuss this issue, before proposing a multi-agent approach to improve NPC adaptability and assessment especially in a collaborative context illustrated with the SIMFOR SG. We then detail the assessment conceptual framework and present briefly the implementation of the new SIMFOR SG as well as some preliminary experimental results. We conclude on the perspectives raised by our contribution.

2 Assessment in Serious Games

Serious Games (SG) are more and more used for training in various domains, but notably in crisis management [5, 8]. Learners' assessment in SG is a recognized as an important research issue [1, 9] and as in real field training exercise, Game-based crisis management assessment often relies on "human" post game debriefing either based on logs analysis, video or interviews [10]. Moreover, it is difficult to produce a collective assessment in this kind of SG, especially when some global goal is shared by all learners but learning outcomes differ from a learner to another. For example, the works presented in [11, 12], provide a framework for serious games and address the concept of learner's assessment and its importance. However, the learner's assessment is performed manually either by the learner itself (self-assessment) or by human monitors. In [7, 13], the authors show the potential of multi-learners in SG and were inspired by MMO RP Gales (*Massively Multi-player Online Role Playing Game*) for designing SG. Research in [7] points the lack of SG for collaborative learning and try to offer a response with *Escape from Wilson Island*, a SG for learning social skills of collaboration. This work falls within the field of Computer Supported Collaborative Learning, CSCL [14]. Nevertheless, the assessment of the collective performance is not addressed, and focuses on the experience of individual game players (with a survey). In [15], team score is computed as the sum of individual scores, and do not take into account players interactions.

It should be noted that these evaluations consider in most cases homogeneous skills as in a participatory teaching in a class of students: whether in a single learning system or group learning, the course of each participant may vary, but the goal of training is unique even when it incorporates the collective dimension. The issue of assessing both individually and collectively heterogeneous skills is not addressed.

The learner assessment can also be approached from the field of Intelligent Tutoring System (ITS) [16] were Learners' assessment is a major theoretical and experimental challenge. Combining the evaluation of an ITS with a serious gaming opportunities, we can thus improve learning outcomes in SG. Among works in ITS, we can cite HAL, Help Agent for Learning [17], an ITS for training TGV drivers and HERA, Help Agent for Learning [4], a training tools for security management in high risk industrial sites. The learner's monitoring and assessment is discussed but remains individual and the issue of collective assessment is not discussed.

Crisis management is a collaborative process that goes through the implementation of several different tasks (depending on the role of the actor and the context) [15], players

have specific educational objectives but share the common goal of managing a crisis. We propose to combine ITS assessment methods to multi-player SG to improve players feedback and thus their learning of collective procedures.

3 How to Improve Assessment in a SG: Illustration with the Simfor SG

SIMFOR (Fig. 1) is a serious game developed by SII company (www.groupe-sii.com) in partnership with Pixxim company (www.pixxim.fr). SIMFOR is a multi-player game training for crisis management by allowing different people to learn skills (shared or specific). Managing a major crisis can mobilize several hundred stakeholders, from the regional Prefect in his office to the firefighter in the field. These stakeholders are required to communicate and work together in order to restore a normal situation. The project objective is to immerse users in a simulated real-time crisis management situation, realistic in terms of environment, self-evolving scenarios and actors (roles). Initially based on a human assessment of the players' skill and simplified NPC (Non-Played Characters), SIMFOR SG can benefit from Distributed Artificial Intelligence by: (i) improving the NPC simulation (complex behaviors and interaction); (ii) guiding the players assessment. Both objectives can be attained with a Multi-Agent Systems (MAS) approach.

Fig. 1. Different game user interfaces of the SIMFOR SG

NPCs are used to adapt the crisis management exercise perimeter to the available stakeholders as well as to specific training objectives. Therefore, SIMFOR is a heterogeneous collaborative learning SG, where tasks are performed by different actors with a common purpose but each one with specific individual objectives. Thus, SIMFOR must deal with two types of learners' assessment: individual and collective. Solving the crisis requires the resolution of all procedures of the stakeholders, so individual evaluation can affect the collective evaluation, and conversely the collective evaluation can affect the individual evaluation too. For example if a learner has successfully executed his procedures, but the main purpose was not reached (material and human loss for example), the learners must be evaluated on their individual and collective performance to infer the reason of failure (lack of communication, missing procedure of another learner, …). The following sections explain how player's immersion and assessment can be improved.

3.1 Agent Oriented Simulation for Realistic Adaptive NPC

The use of MAS to develop software avatars simulating the behaviour of human players is not new [18]. The interest of the MAS for SG is well known (realistic and adaptive behaviour, modularity, behaviour models understandable by non computer science experts, organisational modelling …) [2], so this section will present briefly how NPC simulation capabilities of the SG has been dealt with. To design our agents, we have adopted a BDI architecture (Beliefs, Desires, Intentions), a classic approach to design agents using deliberative behaviours, giving them a certain ability to adapt through complex behaviours [19]. A design tool has also been developed to facilitate the design of game scenario in terms of behaviours and agent types. The NPC must also be able to interact with other players (human or NPC) and act during a game session, depending on the state of the physical environment (represented by the 3D environment in SIMFOR). We propose an *ad hoc* BDI model as a set of agents, actions and facts. The agent model is:

$$\text{Model}_{(\text{role})} = \{\text{Goal},\ \text{Plans},\ \text{Facts},\ \text{Dialogues}\}\,. \tag{1}$$

A game agent (GA) seeks to achieve its goals (assets whose preconditions are verified) by activating the appropriate plans based on its knowledge (defined as a declarative list of facts). Each plan consists of actions directed towards the environment or towards other actors producing effects on this environment or sending messages. The concept of *effect* reflects the social actions and physical influence of agents in the environment. The reasoning of the agents includes (implicitly) decisions and actions. In the case of interactions between players (human or not), possible Dialogues are modelled as a tree where nodes are sentences, each having preconditions to be respected and a list of receiver (all roles, list of roles…). This interaction modelling choice is justified by the need to reference each interaction in relation to the behaviour expected in the crisis management procedure learned.

3.2 Multi-criteria and Distributed Assessment: The Evaluation Space Concept

Integrating evaluation in a serious game involves the use of knowledge, information or data produced or processed continuously until the end of the game. Each information requires a specific manipulation (or a reasoning about these knowledge) to extract evaluation. This section develops a modelling contribution which adds learners' assessment capabilities to the SIMFOR SG taking into account the various nature and origin of information elements required to produce these assessments. Each information element requires special handling (or reasoning on this knowledge) in order to produce an assessment. We define here the evaluation space concept, Sect. 4 will instantiate three evaluation spaces, each one constituting a point of view on the learner's assessment.

A natural way to deal with the complexity of this information management (in the broad sense) is to divide and organize this information into homogeneous groups which can have a dedicated primitive to produce an assessment. The evaluation space concept is part of this approach encompassing all the elements needed to produce assessments, considering the game scenario through different views, each corresponding to a particular assessment objective. An evaluation space is defined as:

$$\text{Evaluation Space} = \{\text{Kw}, \text{I}, \text{M}, \text{AM}\}. \quad (2)$$

- *Knowledge representation model* **(Kw):** There is different kind of knowledge (data, facts, procedures, learner model), each based on a specific modelling paradigm (data modelling, rule based, Bayesian networks …). To ensure homogeneity each space has a set of similar knowledge representation language.
- *Indicators* **(I):** An indicator is a quantitative data that characterizes an evolving situation (an action or consequences of an action) in order to evaluate their status. The use of indicators for the learners' assessment is recurrent in SG [9].
- *Metrics* **(M):** The metrics represents the methods and unit of measure used to exploit the knowledge. It can be used to compare expected results following actors' behavior/decision to their actual doings or to analyze an interaction graph. Thus, the metric quantifies the indicator to compute an assessment.
- *Assessment model* **(AM):** There is different model of assessment, depending on the space and his knowledge representation. The assessment can relate to an action or a procedure or a global assessment. An indicator computation relies on a specific assessment model according to its associated metric.

Thus, an assessment model **AM** can be seen as a utility function (see formula 3) that produces an indicator **I** from a subset of knowledge **Kw** (expressed as a mode of representation) and its associated metric **M**.

$$\text{AM}: \text{Kw} \times \text{M} \rightarrow \text{I}. \quad (3)$$

In the particular case of crisis management, the variety of skills and related knowledge may make difficult the (re)design of a SG. By defining the components of a space evaluation, we seek to guide their characterization independently of the application domain, as well as the identification of skills to be assessed in a serious game. The concept of space and evaluation can facilitate the design process of SG.

4 Applying the Evaluation Spaces to a Crisis Management SG

Adding learners' assessment to SG raises issues in terms of representation, manipulation of knowledge and data acquisition, but also assessment methods (in a mathematical sense). The SIMFOR project presents interesting features by its multi-actors and collaborative nature. Therefore two different kind of assessments are needed, either computed in real time (to show the progress of the player) or at the end of the game:

- *Individual assessment* is a summative assessment that assesses and certifies the learning of the learner at the end of a game scenario.
- *Collective assessment* provides an assessment to the collective performance of the group. It is based on the various communications and interactions between actors (learners and NPC) and thus allows to infer a causal relationship between the missions of the various actors (actor A has failed in its mission because the actor B did not send the correct information to simulated actors).

The overall assessment, which can be determined at the end of a game session, will integrate both individual and collective assessment. Certification of competence or knowledge of the learner can be obtained by aggregation of such assessments. After analyzing the different characteristics of the learners' assessment for crisis management (interactions, behavior, environment) we have defined three different areas of assessment: consequences on the Physical environment, Behavioral and Social abilities. These different *Evaluation Spaces* are described below.

4.1 The Physical Evaluation Space

The *Physical Evaluation Space* represents the view of the SIMFOR virtual environment and allow to assess the learners' outcomes from the virtual environment. The knowledge **(Kw)** in this space is based on the information of the 3D models (avatars, transport means, disaster, vegetation, trees, etc.) as well as the meta data from the geographic information system (GIS) such as building type (commercial, residential, school, etc.), the number of people in a building, etc.

The indicators **(I)** used in this space will be used to produce an assessment of both individual and collective assessment. For example, an avatar moving from point A to point B (individual assessments) requires the starting position, the ending position, and the elapsed time. On the other hand the physical space can also provide indicators for collective evaluation: e.g. human and material losses. This indicator can give an idea about the overall performance of the group.

The metric **(M)** allow to quantify (unit, distance calculation function, etc.) the indicators previously defined. For example, the metric for material losses indicator may be the cost in Euros.

The assessment model specify how the learners' performance is computed (e.g. preferred travel time rather than travel cost). The assessment model is represented by an objective function that includes various indicators (human and material losses, means used, etc.) to compute the learners' performance (score).

4.2 Behavioral Evaluation Space

The *Behavioral Evaluation Space* represents the procedural view of SIMFOR and allows the SG to assess the learners' behaviors related to the crisis management.

The knowledge representation in the behavioral space includes the learners' actions and knowledge as well as the different information on the skills and procedures to learn (corresponding to the learner model and the domain model of an ITS [16]). The knowledge **(Kw)** involved in this space is modeled as a set of actions and missions. Each role (assigned to an actor) is assigned a set missions to perform for a given scenario. A mission is composed of a set of preconditions, a set of previous missions, and finally a set of actions to perform to reach the goal of the mission. Actions are declined in different forms depending on the actions that players can perform in SIMFOR such as *phone action*, *fax*, *radio*, *talk*, *move* (walking or with a vehicle), and *daybook* (simulate a web blog for disaster monitoring).

Given the heterogeneous nature of the actions, a specific indicator **(I)** is defined for each action, to compute the action efficiency. For assessing learners' missions, we have identified six general indicators **(I)** measuring if a player has acted accordingly to the situation: such as if he has respected the precondition of the mission and the actions sequence, how well and how fast each action has been executed, idle time etc. (see [20] for detail on the indicators computation).

The metrics used in this ES are mainly represented by the time or logical properties as precondition, or scheduling. To have an equal weight between all indicators, the maximum score for each indicator has been standardized as ranging from 0 to 1.

The assessment model **AM** allow to compute the mission score and it is the average of the scores produced by different indicators defined in Table 1. Each SIMFOR action is assigned a specific assessment model (given the heterogeneous aspect of the actions). For example, for the *phone* action, the indicators (I) are the communication time, the target actor and the information exchanged (in case of textual dialogue with a NPC). The action efficiency is the average of these indicators.

Table 1. Actions performed by the actor for the mission Inform the authorities.

Actions	Reference action	Indicator	Score
Phone(duration, target, msg) = (77, officer, {TMD})	Phone(45, officer, {TMD})	ActionPhoneIndicator(stime, target, msg) = (45/77, 1, 1)	0.86
Fax(duration, target, faxName) = (232, officer, informationSheet)	Fax(120, officer, informationSheet)	ActionFaxIndicator(stime, tatget, faxName) = (120/132, 1, 1)	0.83
Fax(duration, target, faxName) = (13, mayor, informationSheet)	Fax(120, mayor, informationSheet)	ActionFaxIndicator(stime, tatget, faxName) = (1, 1, 1)	1
…	…	…	…

In order to illustrate the learners' assessment in the behavioral space, we'll take a sequence of game play for an actor and try to analyze the influence of different indicators. Here, the CODIS actor (Departmental Center for Operational Fire and Rescue Services) has received information about the accident from a firefighter on the scene. The CODIS mission is to *inform the authorities* of the TDM accident.

Actions to achieve this mission are described in Table 2. The first column shows the action performed by the learner with the call duration, the target actor, and the messages exchanged. The second column shows the reference action (from the domain model). The third column represents the defined indicator to assess the action *phone*, the score

will be an average between the time score (ratio between expected and real duration of the action), a score related to the target actor (0 or 1 if the right actor is contacted), and a score related to the messages exchanged (set of sentences exchanged during the dialogue). The execution time of the last action is low because the learner has sent the same fax to the officer, and only has to retransmit it.

Table 2. Mission assessment: inform the authorities (CODIS)

Indicators	Outcome	Score
Precondition	*Prec = {TDM}*	1
Order	*Actions = {phone, fax, fax, fax, fax}*	1
Actions count	*Nb action = 5, nb actions (expert) = 5*	1
Duration	*Time = 589, estimated time = 525*	0.7
Idle time	*Idle time = 246*	0.03
Actions efficiency	*Action efficiency = {0.86, 0.83,1, 1, 1}*	0.93
Mission score		0.77

Table 2, describes how the indicators of the mission and the scores produce the mission assessment. The first indicator is the respect preconditions, the learner (CODIS role) triggered the mission after receiving information *(TDM)* from the firefighter on the scene and thus satisfies the precondition *TDM* and the score is equal to 1. The second and third indicator relates to the sequence of actions the learner has to perform (correct number) and in the correct order (both scores are 1). The execution time of the mission is the fourth indicator. The difference between the duration provided by the expert and the player actions are compared relatively to the duration itself (to differentiate small variations compared to short or long duration). The fifth indicator relates the idle time of the learner, calculated by accumulating the idle time between each action of the mission (inactivity score is 0.03). Idle time has a great influence on the score for crisis management, the learner must not lose time between the executions of each action of the same mission.

The last indicator relates to the actions efficiency, this indicator is calculated by the average of the actions efficiency scores presented in Table 3. The overall score of the mission is the average of six indicators and is equal to 0.77. After (or during) the game, the learner can have information on its performance, and can see that his idle score is low and he was not reactive enough for executing the mission actions.

Table 3. Exercises result from the TDM scenario.

Learners	Exercise 1 -				Exercise 2			
	Individual	*Social*	*Missions*	*Global*	*Individual*	*Social*	*Missions*	*Global*
CODIS	0.817268	0.804907	0.82963		0.85349	0.845879	0.861111	
Mayor	0.943310	0.882632	1		0.957254	0.914508	1	
Prefect	0.755899	0.865317	0.646481		0.763133	0.852076	0.675990	
Sub-Prefect	0.763071	0.88662	0.639522		0.783950	0.889330	0.678570	
Group		0.859869	0.778090	0,65		0.875448	0.803918	0,76
Global				0.762653				0.813122

4.3 The Social Evaluation Space

The *Social Evaluation Space* represents the interaction (simple communication, coordination/cooperation …) between different actors and including the collaborative dimension of learning. The social space is represented by a social graph that describe every interaction between actors and allows to compute an interaction strength between each actor as well as the global coupling of network. Assessment will be based on these measures. In [2], we find considerations on the representation and exploitation of interactions for the learners' assessment. The indicators presented in this space are the network coupling and the strength of interaction between actors and are calculated as in [21]. Short and frequent exchanges (between actors) produce a strong coupling, while long and rare exchanges reflect a weak coupling. This indicator will determine if the actor has interacted with the right actors.

The Knowledge **(Kw)** modeled in this space is represented by a set of interactions, wherein each interaction is characterized by:

- Actor at the origin of the interaction.
- Actor target of the interaction.
- Interaction type (phone, fax, radio, …).
- Interaction date.
- Interaction duration.

Indicators in the social space are based on the network coupling. The interactions between the actors enable us to compute the network coupling like the interaction network in [22], but taking into account specific considerations to the field of risk management and more particularly the study of interactions between firefighters in operation [23]. The metric used to compute the network coupling are based on the number of interactions and their duration, and the time between each interaction. From the coupling between actors, we compare the coupling of the exercise and the coupling provided by the domain expert (calculated on the domain model). If the coupling is weak between A and B whereas strong in the reference graph, it means that there was a lack of communication between A and B during the exercise.

Figure 2 shows an example of a network coupling of SIMFOR exercise. In graph (a), the maximum coupling is obtained by the relationship between the prefect and subprefect. In graph (c), the interaction between the prefect and the mayor is red (dashed line), showing thus a negative difference (the coupling during exercise between the two

actors is lower than the reference coupling revealing a lack of communication). The interaction between CODIS and fireman is blue (dotted line), showing a positive difference (coupling during exercise between the two actors is greater than the reference coupling i.e. surplus of communication). Thus, by computing the network coupling, we can address the individual assessment of an actor in his relationship with others (comparing interactions between actors) and the collective assessment (through the global network coupling) by comparing the result of the exercise with the coupling provided by the expert (extracted from the domain model).

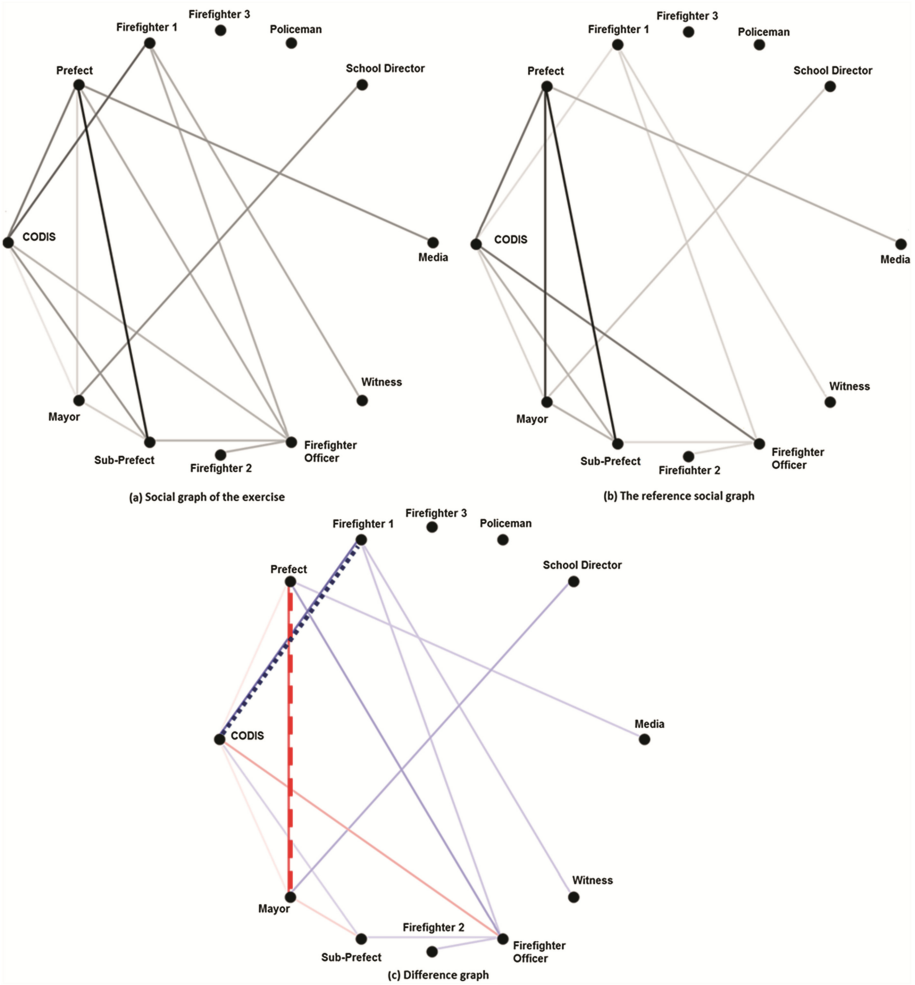

Fig. 2. An example of the network coupling for a SIMFOR exercise. The graph (a) shows the interaction force of the group during the exercise (a black color correspond to the maximum coupling). The graph (b) shows the reference graph computed from the domain model defined by the expert. The graph (c) shows the difference (b-a) between the graph (a) and the graph (b). Thicker lines indicate strong coupling and conversely thin lines low coupling (Color figure online).

5 Implementation and First Experiments

This section introduces the general architecture of our system and presents some preliminary experiments results.

5.1 General Architecture

The SIMFOR architecture combines elements from the Intelligent Tutoring System and Serious Game domains (Fig. 3). Our goal is to associate the playful learning of SG and the different modules of an ITS (domain model, learner model, pedagogical model) to get the optimal learning environment. The SIMFOR architecture is composed of the following components:

- **The SG module (SIMFOR):** this module includes the 3D models, user interface (as a communication channel between the learner and the system), simulation module (for natural phenomena such as fire propagation), and data models. This module constitutes the former "perimeter" of the SIMFOR SG to which behaviour simulation and user assessment capabilities are added.
- **The Behaviours Simulation module:** allows simulating humans' behaviours to replace absent players with "artificial" actors (Game Agent).
- **The Evaluation module:** the evaluation module provides skills assessment of players in real time to the pedagogical module.
- **The Pedagogical module:** which plays the role of a virtual tutor accompanying the learners by providing support and help during (and after) their training.
- **Knowledge representation module:** All knowledge used or produced by the previous modules of our proposed architecture are stored as an ontology in the domain model and the learner model. The ontology describe the general domain of crisis management (adapted to the SIMFOR context). The *Domain model* represents the general concepts of crisis management and is segmented into parts representing a role or a skill to learn. For each learner or agent, a *Learner Model* is associated, which represents its mental state at a time *t*.

The SIMFOR SG and its additional modules has been developed in C++ with QT interfaces. The evaluation module is based on agents Detail on the multi-agent architecture is described in [20]. Briefly, These agents collect learners' data, process and evaluate learners data and provide support to learners. They also simulate actors by acting in the 3D environment and exchanging messages with other actors.

5.2 Experimental Result

In this section we present an example of crisis management scenario, the case of a Transport of Dangerous Materials (TDM) scenario which is interesting as it involves several roles such as fire-fighters, police, mayor, prefect … Moreover, the TDM scenario may evolve into an environmental pollution scenario (chemicals leakage), or large fire disaster (flammable products), if badly handled.

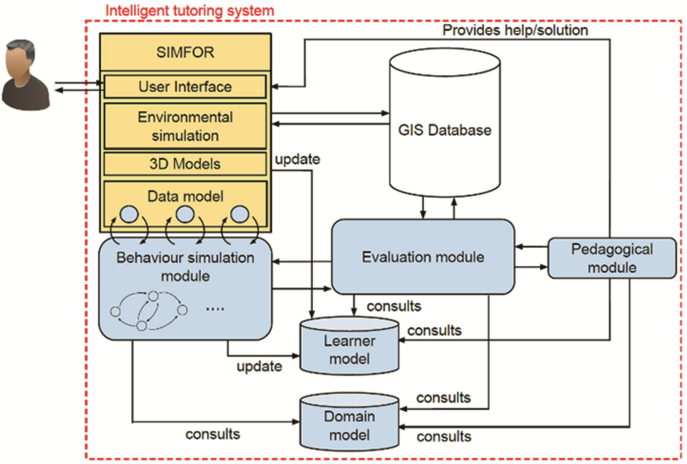

Fig. 3. The general architecture of SIMFOR

The TDM scenario was developed with the help of a crisis management expert and aims to sensitize the stakeholders to the different consequences that may result from a TDM accident. The scenario begins with a TDM truck overturned after a traffic accident on a roundabout in the outskirts of Arles (city in south of France), close to a school, the Rhone river and a railway. The tank is damaged and hydrocarbon spills on the road. A witness to the accident gave the alert. The domain expert defines beforehand (via the domain model editor) the missions related to each role as well as the exercise scripting (via scripts that can trigger events under certain conditions).

To start the exercise, we must first run a SIMFOR server with the selected scenario exercise. Once the server is launched, players connect to the server, either by internet or local network (for our tests, the exercise was a local network but learners were in different rooms). Once all learners are connected the exercise starts. In our first experiment, four learners and seven NPC are engaged in the TDM exercise.

The TDM scenario begins with the CODIS sending a fireman at the accident scene to interrogate the witness and gather information about the accident. Once the information confirms a TDM accident, the CODIS warns an officer (fireman) and drafts an information sheet on the incident and sends it by fax to the officer, the mayor, prefect and sub-prefect. The officer, for his part, must send a second fireman on the disaster scene with a Water-tender vehicles and gives first step instructions to the fireman in scene. Once the information sheet received, the prefect must discuss with the sub-prefect to agree on the location of the forward command post (FCP) and triggers ORSEC plan (filled and distributed via fax by the prefect). The civil security advise their actions through the daybook tool after each major action. The sub-prefect is responsible thereafter to inform the CODIS and the police of the FCP place before he goes, and discuss with the mayor to identify the potential risks around the disaster (school evacuation, area risk, etc.). The officer uses the FCP truck to go to the FCP place and engage in a debriefing with the fireman on the progress of the situation and reports it to the sub-prefect at FCP. Depending on the severity of the situation (pollution, hydrocarbon fire),

the sub-prefect broadcasts a press communication to the media and requests an additional resources if needed. If the situation requires, the sub-prefect should contact the mayor to evacuate nearest building of the disaster such as the school. Once the disaster controlled, the prefect sends a prefectural order (decree), to call the end the ORSEC plan, to the CODIS, mayor and policeman.

All evaluation results and the learners' feedback are saved in a XML file containing for each learner:

- List of performed actions.
- Assessment result for each action performed.
- The history feedback of the learner.
- The assessment of each mission completed.
- The social and the final score of the learner.

In addition to individual assessments, the resulting file contains the history of all interactions between the different actors, which allows as to compute the interaction strength (network coupling), the difference graph and social score to produce a global and collective assessment (with the integration of other global indicators depending on the scenario exercise). Additionally, the result file provides also other indicators purely observable such as the global coupling history (we can deduce the key moments of the scenario exercise due to spikes in the coupling graph), or the actions sequence on the time scale.

This experiment aimed to (i) check the automatic assessment relevancy and (ii) to evaluate the learners' progress using the SIMFOR SG. To assess the first point, learners receive question forms related to the scenario exercise in order to compare these results to the automatic assessment provided by SIMOR. For the second point, we have performed two exercises and compared the learners' performance evolution between the exercises. The scores obtained are exposed in Table 3.

The exercise results in Table 3 show that all learners have improved their performance between the two exercises. For each exercise, the first column contains the players' final individual score as the average of their individual behavioural scores (i.e. how well they have executed their mission) and social. The "group" row lists the collective (as a group) social score (average of all individual social scores), behavioural score (average individual behavioural/mission score), and physical score. The collective physical score is related to the area finally affected by the disaster which grows up until the disaster is controlled (fire propagation simulation). The last row gives the general performance of the group. All scores computed are normalized to the interval [0, 1] (actions, missions, social, physic, individual and collective).

Table 4 presents the players' score resulting from their answers to the question form (with specific questions related to their role). The forms' scores are coherent with the automatic assessment: all learners have improved their performance, and the consistency of the automatic assessment compared to the forms results is confirmed except for the prefect and the sub-prefect due to many and long procedures to do introducing some noise in the assessment (additional actions, delay, …), but combining the missions score with the social score refines the automatic assessment.

Table 4. Forms results from the TDM scenario.

Learners	Exercise 1		Exercise 2	
	Forms responses	*Mission & Individual scores*	*Forms responses*	*Mission & Individual scores*
CODIS	60%	0.82963 - 0.817268	91%	0.861111 - 0.853495
Mayor	100%	1 - 0.943310	100%	1 - 0.957254
Prefect	81%	0.646481 - 0.755899	81%	0.675990 - 0.763133
Sub-prefect	75%	0.639522 - 0.763071	87%	0.678570 - 0.783950

6 Conclusion

With the growing interest of serious games for training, the issue of learners' assessment is increasingly crucial. This paper has presented how to improve NPC simulation and players assessment in a Crisis Management SG with an adapted BDI model and the Evaluation Space framework. Assessment is realized through a multi-criteria (evaluation of different trades and skills) and distributed (via dedicated evaluation spaces) assessment system supported by a multi-agent system. The new SIMFOR SG has been implemented and tested on a realistic TDM scenario, but more extensive on-field experimentation is required for complete validation.

Future work may consider adding emotional factors in the NCP behaviour (simulating panic) which can simulate time/stress constraints to the players facing such behaviours. Collective assessment is promising and deserves more investigation. The Evaluation Space concept allows quantitative evaluation of the interactions, but can be extended to qualitative interaction by analysing the content of the interaction (and not only who and when), but would not call into question the present framework as only new low level agents is required, confirming the relevancy of our approach.

References

1. Nieborg, D.: America's Army: More than a game. Transforming Knowledge into Action through Gaming and Simulation, SAGSAGA (2004)
2. Mathieu, P., Panzoli, D., Picault, S.: Virtual customers in a multiagent training application. In: Pan, Z., Cheok, A.D., Müller, W., Liarokapis, F. (eds.) Transactions on Edutainment IX. LNCS, vol. 7544, pp. 97–114. Springer, Heidelberg (2013)
3. Haferkamp, N., Kraemer, N.C., Linehan, C., Schembri, M.: Training disaster communication by means of serious games in virtual environments. Entertainment Comput. **2**, 81–88 (2011)
4. Amokrane, K., Lourdeaux, D., Burkhardt, J.M.: Hera: learner tracking in a virtual environment. IJVR **7**(3), 23–30 (2008)
5. Di Loreto, I., Divitini, M.: Games for learning cooperation at work: the case of crisis preparedness. In: ECTEL-meets-ECSCW, pp. 20–24 (2013)
6. Buche, C., Querrec, R., De Loor, P., Chevaillier, P.: Mascaret: pedagogical multi-agents systems for virtual environment for training. In: International Conference on Cyberworlds, pp. 423–430 (2003)
7. Wendel V., Gutjahr, M., Göbel S., Steinmetz, R.: Designing collaborative multiplayer serious games for collaborative learning. In: Proceedings of the CSEDU (2012)
8. GALA Consortium: Learning Analytics for SGs, Deliverable 2.4. Technical report (2014)

9. Thomas, P., Labat, J.-M., Muratet, M., Yessad, A.: How to evaluate competencies in game-based learning systems automatically? In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 168–173. Springer, Heidelberg (2012)
10. van Ruijven, T., Mayer, I., de Bruijne, M.: Multidisciplinary coordination of on scene command teams in virtual emergency exercises. IJCIP **9**, 13–23 (2015). Elsevier
11. De Freitas S., Jarvis S.: A framework for developing serious games to meet learner needs. In: The Interservice/Industry Training, Simulation & Education Conference. NTSA (2006)
12. Yusoff, A., Crowder, R., Gilbert, L., Wills, G.: A conceptual framework for serious games. In: 9th Conference on Advanced Learning Technologies, ICALT 2009. IEEE (2009)
13. Raybourn, E.-M.: Adaptive thinking and leadership training for cultural awareness and communication competence. Interact. Technol. Smart Educ. **2**(2), 131–134 (2005)
14. Stahl, G., Koschmann, T., Suthers, D.: Computer-supported collaborative learning: An historical perspective. In: Sawyer, R.K. (ed.) Cambridge Handbook of the Learning Sciences. Cambridge University Press, Cambridge (2006)
15. Oliveira, V., Coelhoa, A., Guimarães, R., Rebelo, C.: Serious game in security: a solution for security trainees. In: VS-GAMES 2012 (2012)
16. Burns, H., Capps, C.: Foundations of intelligent tutoring systems : an introduction. In: Polson, M.C., Richardson, J.J. (eds.) Foundations of intelligent tutoring systems, pp. 1–18. Lawrence Erlbaum Associates, Hillsdale (1989)
17. Lourdeaux, D., Burkhardt, J.M., Bernard, F., Fuchs, P.: Relevance of an intelligent agent for virtual reality training. Int. J. Continuous Eng. Life-long Learn. **12**(1/2/3/4), 131–143 (2002)
18. Chang, P.H.-M., Chen, K.-T., Chien, Y.-H., Kao, E.C.-C., Soo, V.-W.: From reality to mind: a cognitive middle layer of environment concepts for believable agents. In: Weyns, D., Van Dyke Parunak, H., Michel, F. (eds.) E4MAS 2004. LNCS (LNAI), vol. 3374, pp. 57–73. Springer, Heidelberg (2005)
19. Rao, A., Georgeff, M.-P.: BDI agents: from theory to practice. In: ICMAS 1995 (1995)
20. Oulhaci, A., Tranvouez, E., Fournier, S., Espinasse, B.: A multi-agent system for learner assessment in serious games: application to learning processes in crisis management. In: Seventh IEEE International Conference on Research Challenges in Information Science, Paris (2013)
21. Miller, J.-G.: Living Systems. Mcgraw-Hill, New York (1978)
22. Kay, J., Maisonneuve, N., Yacef, K., Reimann, P.: The big five and visualisations of team work activity. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 197–206. Springer, Heidelberg (2006)
23. Baumard, P., Vidal, R.: Fiabiliser la gestion des feux de très grande ampleur - enhancing reliability in large scale willand fire response organization. Ministère de l'écologie, de l'énergie, du développement durable et de la mer. Technical report (2009)