Using AI Techniques in a Serious Game for Socio-Moral Reasoning Development

Tato Ange,[1,2,4] **Nkambou Roger,**[1,2] **Dufresne Aude**[3]
[1]Department of Computer Science, Université du Québec À Montréal, Canada
[2]Centre de Recherche en Intelligence Artificielle, Montréal, Canada
[3]Department of Communication, Université de Montréal, Canada
[4]nyamen_tato.ange_adrienne@courrier.uqam.ca

Abstract

We present a serious game designed to help players/learners develop socio-moral reasoning (SMR) maturity. It is based on an existing computerized task that was converted into a game to improve the motivation of learners. The learner model is computed using a hybrid deep learning architecture, and adaptation rules are provided by both human experts and machine learning techniques. We conducted some experiments with two versions of the game (the initial version and the adaptive version with AI-Based learner modeling). The results show that the adaptive version provides significant better results in terms of learning gain.

Socio-Moral Reasoning (SMR) is a socio-cognitive construct essential for decision-making, as well as social interaction adaptation. It is commonly defined as ”how individuals think about moral emotions and conventions that govern social interactions in their everyday lives” (Beauchamp, Dooley, and Anderson 2013). Being able to predict and diagnose one's socio-moral reasoning skill level (or ability) is a key step for quantifying peoples' social functioning and can be used to identify those at risk for mal-adaptive social behavior. This diagnosis could help orient people towards appropriate services or provide adequate support to improve this skill's development. The Socio-Moral Reasoning Aptitude Level (So-Moral)(Dooley, Beauchamp, and Anderson 2010) task is a computer-measured walkthrough in which children and adolescents are presented with visual social dilemmas from everyday life. They are then asked to verbalize how they would react in this situation, justifying their answer. The participants' answers are recorded verbatim in transcripts that are subsequently scored manually by experts using a moral-maturity coding scheme inspired by Kohlberg's theory of moral development (Kohlberg 1984). Verbatims are short or long text containing at least one sentence. Each socio-moral reasoning level was well documented by experts.

In this paper, we present a novel emotionally adaptive serious game (LesDilemmes) designed to help develop SMR maturity. The game is based on the existing, computerized So-Moral task. The user model is build using a novel hybrid neural network architecture and the adaptation rules are provided by both human experts and machine learning techniques (a decision tree and a neural network). We will show that the game can significantly improve the SMR of teenagers.

LesDilemmes : A serious game to learn SMR

Many studies have proven the pedagogical value of serious games (Dondlinger 2007; Andrews 2011), which can help build concrete spaces where abstract problems can be explained. For serious games to be efficient it is important that they rely on knowledge models and good learner models. Playing serious games can improve diverse cognitive abilities (Granic, Lobel, and Engels 2014), and can provide opportunities for superior learning experiences since the learner is active and can experience different situations that reflect behaviorally and decision making capacities (Ryan and Deci 2008). Even though there exist several serious game in different domain, there are none that target the SMR skills.

LesDilemmes aims to assess and improve the socio-moral reasoning skills. The game involves a series of dilemmas that the player must solve. It has been implemented as a 3D environment to recreate as faithfully as possible, the real situation in which the individual would normally have to make decisions. For each dilemma, the player has to decide what action to take and give a verbal justification of his choice. The experts then use the reasons given by the player to evaluate his level of maturity. This evaluation is done automatically in the game thanks to our hybrid model that we will present later in this paper. The level of maturity (which varies between 1 and 5) is determined according to the content of the individual's verbatim and the information extracted from a coding system (see table 1) (Chiasson et al. 2017). The player is also surrounded by non-player characters (NPCs) who represent different SMR levels. He can consult the NPCs and evaluate the justifications they suggest for each dilemma. The player's emotions are captured during the game, using the Facereader[1] tool (using facial ex-

[1] shorturl.at/hlCN4

Table 1: Brief description of So-Moral coding and examples (Chiasson et al. 2017).

Level	Brief description	Example
1	Moral justifications have an egocentric focus, which is based on obedience to higher authorities and potential consequences to themselves for their actions (e.g. punishment). Thinking at this level is inflexible; there is only one right/wrong way to act.	Because I could go to jail.
2	Moral justifications are based on a concept of pragmatic deals or exchanging favors with others ('fair deals'). Thinking is more flexible and is determined by context. The correct option is the one that is right for oneself (self-interest).	Because i might need his/her help in the future.
3	Moral justifications have a focus on interpersonal relationships, a sense of 'goodness', and feelings such as empathy and trust. Decisions are made with good motives and a prosocial perspective of the world.	Because he/she could get hurt.
4	Moral justifications start to incorporate a broader view of morality; based on the compliance with rules, regulations and standards that society has established to ensure social order	Because if everyone were to be unfaithful, relationships would not have any meaning.
5	Moral justifications are characterized by the capacity to evaluate situations from various points of view to identify values involved in the specific situation to make the fairest decision. Protection of fundamental values and people's rights is specific to this stage, even though these concepts are expressed very concrete.	Because people work hard for their things and we should respect their belonging.

pressions).

The learner model

The learner's model has two parts: a model to predict the SMR level, and a model representing the emotional state.

Emotional state

The player's emotions are represented by a declarative memory (key-value list) where the keys represent emotions. We considered the 7 basic emotions of Ekman (Ekman 1999), namely: Neutral, Happy, Sad, Angry, Suprised, Scared and Disgusted, plus the Valence and the Arousal. The latter two are calculated according to well-defined formulas, and all values range from 0 to 1 except the valence which ranges from -1 to 1. The emotions are calculated in real-time using the Facereader tool.

Knowledge state

The knowledge state in LesDilemmes represents the SMR level of the player. Automatically assessing an individual's SMR maturity level requires to analyze in real-time with appropriate solutions, verbatims provided when solving dilemmas. We have a set of textual data (verbatims) already annotated by experts. These annotated data are accompanied by an associated description (a paragraph with key concepts) of each level of SMR (see table 1). We have developed a hybrid model that allows us to accurately assess a player's SMR level based on his justification (verbatim).

Hybrid models Towel et al. (Towell and Shavlik 1994) defined hybrid learning techniques as ”methods that use theoretical knowledge of a domain and a set of classified examples to develop a method for accurately classifying examples not seen during training”. Therefore, a hybrid learning system should learn more effectively than systems that make use of only one of the information sources. There exists few works that focus on the combination of the a priori knowledge and deep learning architectures. Among them, Towel et al. (Towell and Shavlik 1994) (which might be the first paper to discuss this matter) proposed a hybrid system called KBANN (Knowledge-Based Artificial Neural Networks). Their solution maps expert knowledge, represented in propositional logic, into neural networks and then refines this reformulated knowledge using back-propagation. Coro et al. (Coro, Pagano, and Ellenbroek 2013) combined Neural networks (NN) with simulated expert knowledge. The simulated expert was used to generate some examples, which were added to the training set of the NN.

In education, a priori expert knowledge is usually available and generally used to build Intelligent Tutoring Systems (ITS). In other domains, expert knowledge can be available through books or previously built models (such as rules-based models). We believe that this a priori expert knowledge, sometimes acquired over decades of intense research, cannot be dismissed and ignored. In the present paper, in particular, we put forth an approach that uses the attention mechanism (Luong, Pham, and Manning 2015) and capitalizes on the availability of (possibly simplified or inaccurate) theoretical models to reduce the amount of empirical data to use. To our knowledge no research has proposed to combine a priori knowledge with deep learning architecture using the attention mechanism. Moreover, no research in the educational data-mining domain has yet focused on this matter despite the availability of expert knowledge. We applied the proposed solution to the automatic detection of SMR skill level of learners in LesDilemmes.

Automatic prediction of SMR level We used two different NLP (Natural Language Processing) techniques to make the descriptions (knowledge) usable by the hybrid model. The first technique is the *Word Movers' Distance* (WMD) (Kusner et al. 2015) and the second is the n-grams (Damashek 1995). WMD is a technique that allows to submit a request and return the most relevant documents to the

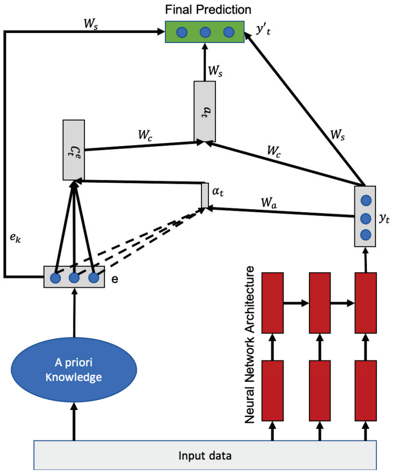


Figure 1: Global attentional hybrid model

request. The purpose of using these techniques is to be able to compare the input data (verbatim) with the knowledge and to output a result that will, therefore, be combined with features extracted by the neural architecture using the attention mechanism, .

The proposed hybrid model will constrain the deep neural network model (DNN) to pay attention to what the a priori knowledge says about the current input x. Since attention (Xu et al. 2015) is a memory-access mechanism, it fits well in this context where we want the model to have access to the a priori knowledge during learning. In other words, the DNN will ”consult” the knowledge before taking the final decision. The importance the DNN model will accord to what the knowledge says is computed (learned) through attentional weights (W_a and W_c) (see figure 1). As the training goes, the neural model will know the importance it should give to each of the predictions from the knowledge. W_a corresponds to the weights calculating the importance of each feature learned by the neural architecture (y_t) with respect to each feature extracted from the knowledge. W_c corresponds to the weights measuring the importance of predictions made from knowledge (via the context vector) and learned characteristics (y_t) for the estimation of the final prediction vector (see figure 1).

Thus, the model will focus on what the knowledge says before taking any decisions. In the attention mechanism presented by Luong et al (Luong, Pham, and Manning 2015) (specifically the global attention model), the attention vector is calculated from the target hidden state h_t and the input hidden state. Instead of the hidden input state ($\bar{h}_s$), we will have the data from the knowledge, that will be used to calculate the context vector C_t (which we will call context vector expert side) (see Figure 2 in (Luong, Pham, and Manning 2015)). Thus, given the hidden state y_t (the prediction) of the neural model, and the context vector on the expert side c^e_t, we use a concatenation layer to combine the information of the two vectors to produce the attention hidden state a_t as follows:

$$a_t = tanh(W_c[c^e_t; y_t]) \tag{1}$$

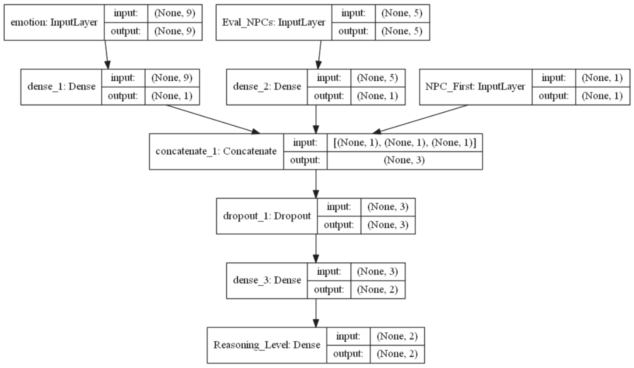


Figure 2: The multimodal neural network for the extraction of adaptation rules.

The attentional vector a_t as well as the prediction made from the expert knowledge e and the hidden state y_t (the prediction) of the neural model are then sent in a dense layer to produce the expected result y'_t. The context vector on the expert side is then calculated as follows:

$$\begin{aligned} score(e_k, y_t) &= e_k \cdot y_t \cdot W_a + b \\ \alpha_{t,k} &= \frac{exp^{score(e_k, y_t)}}{\sum_{j=1}^{s} exp^{score(e_j, y_t)}} \\ c^e_t &= \sum_k \alpha_{t,k} \cdot e \end{aligned} \tag{2}$$

Where $1 <= k <= s$, e is the prediction made from expert knowledge, y_t is the current prediction made by the neural architecture and s is the size of the predicted vector (the number of classes to predict). e is a vector of length equal to the size of the vector to be predicted where each input represents the probability that the input belongs to each class according to the a priori knowledge. e_k is size 1 and W_a,W_c,W_s,W_s,W_s,y_t,e and a_t are of size s. The score is a content-based vector that calculates the correlation (alignment score) between knowledge and latent *features* learned by neural architecture. This parameter defines how expert knowledge and *features* learned from the data are aligned. The model assigns a score of $\alpha_{t,k}$ to the pair of entities at position t and the knowledge (e_k, y_t), based on their correspondence. The set of $\alpha_{t,k}$ are weights defining to what extent each feature of the data from the a priori knowledge must be taken into account for each output (final prediction). The figure 1 shows in detail this global process.

We used the WMD to construct the first part of the vector representing the expert knowledge in our hybrid model. Each description of each level of SMR and each verbatim are considered as different ”documents”. We transformed each document into a representation using a french word2vec pre-trained vectors [2]. We calculated the similarity between each verbatim and each description (5 descriptions) using the gensim tool Wmd-Similarity [3]. For each verbatim, this

[2] http://fauconnier.github.io/

[3] shorturl.at/aejsy

calculation provided us with a vector of length 5 where each entry is the similarity between the verbatim and the description of the corresponding SMR level. We used the second technique to extract the n-grams (uni and bi-grams) from the textual description of the levels, which gave us a list of n-grams for each SMR level. We also generated a list of synonyms for all keywords (extracted manually) included in the descriptions. For each verbatim, we counted the number of times each n-gram of each level appeared in the text. The sum of the vectors generated by this process gave us, for each verbatim, a vector of size 5 where each entry represents the number of n-grams and the number of synonyms of each level found in the verbatim. We finally applied the softmax function to the resulting vector that was added to the one generated by the WMD method. The result vector v of this step is the ”knowledge” vector of our hybrid model. The code for this model is publicly available[4].

The Adaptation model

To introduce a form of feedback and scoring into the game, we have added simulated social feedbacks, showing the number of *friends* and *likes* according to the players' SMR level. When the SMR level of the player's increases, the player gains *likes*, and when he positively evaluates (in agreement) the opinions of NPCs with a higher level of maturity than his own, he gains *friends*. On the other hand, if he agrees with NPCs with lower SMR levels, it does not affect the number of *likes* or the number of *friends*. These rules, are a result of interactions with experts.

Rules defined by experts

Here are some rules defined by the experts:

- **If** the player (with a SMR level equals to R) does not agree with a NPC with a SMR level equals to R_x and $R_x >= R$ **Then** the NPC will display a negative emotion with a thumb down to indicate to the player that it should not disagree.
- **If** the player's SMR level is improving **Then** the feedback is a congratulation message specific to his current SMR level.
- **If** the player has a SMR level equals to 1 **Then** the feedback is a message that will help him think like someone with a SMR level equals to 2 or 3 (the choice of messages is automatic). If the SMR level equals to 2 (selfish), **then** the feedback is a message that will help him think like someone with a SMR level equals to 3 or 4 etc.

Rules extracted from data

We extracted some rules using a decision tree and a neural network. The goal is to automatically determine the attributes of learners and the elements of the system that have an impact on the learning process. Once this step is completed, the actions to be taken (system elements to be modified) should be automatically determined when certain values for these attributes are observed.

We have built a decision tree from the data collected during the first experiment with the non-adaptive version of the game. The class to predict was the SMR level, based on the 7 emotions plus the valence and the arousal, the visiting style of the NPCs and the evaluation done on each of the opinions of the NPCs. The disadvantage of such a solution is that if the data is very large, it becomes less efficient to build a decision tree.

In addition to the decision tree, we have developed a neural network (NN) for the extraction of other rules. Many approaches have been proposed for extracting rules from neural networks. Essentially, rule extraction algorithms fall into three categories: decompositional, pedagogical and elective (Bologna and Hayashi 2018). We used the decompositional approach, where rules are extracted at the level of hidden neurons and output neurons by analyzing the weight values (Murdoch and Szlam 2017). A NN implements a function that takes inputs and produces predictions. Each input has a certain importance in predicting the output. This importance is measured by the weight assigned by the NN to each of the inputs. For example, if we consider a single-layer NN, with 2 neurons on the input layer, then the output would be written as follows:

$$Y = f(w_1 \cdot x_1 + w_2 \cdot x_2) + b \qquad (3)$$

Where f is an activation function, b is the bias and w_i are the weights learned by the network which measure the importance of each input on the output prediction. For example if $w_1 > w_2$ then this implies that the input x_1 is more important in the calculation of y than x_2. This is true if we force the network to learn only positive weights and transform the inputs so that their domain of definition is the same. It has been proven that by constraining the value of weights in this way, the learning process of the network is not affected (Chorowski and Zurada 2015). From our example, we can extract a rule that says: if x_1 is large then so is Y (knowing that f is a strictly increasing activation function on $\mathbb{R}$). To extract the rules from the NN (see figure 2) trained on the same data as the decision tree, we used the following process. Let W be the weight matrix of the penultimate layer where each row $W_j = [w_1, ..., w_n]$ represents the weights connected to each neuron j of the last layer. n = 3 (emotions, evaluation of NPCs, NPCs) and j $\in$ 1.2 where $j = 1$ means that the weights are related to the neuron that triggers a value close to 1 when the reasoning level is less than 3 ($[1, 0]$). To assess the importance of inputs in predicting output, we did not directly consider the weight matrix, but the relative values. So instead of using w_i we used a_i which is the result of applying the softmax function to all weights connected to the same neuron.

$$a_{i,j} = \frac{e^{w_{i,j}}}{\sum_{i=1}^{n} w_{i,j}} \qquad (4)$$

The value $A_{i,j}$ used to evaluate the importance of the i input (in the prediction of each of the j outputs) is calculated as follows:

$$A_{i,j} = \frac{\sum_k a_{i,j,k}}{\sum_k k \sum_j a_{i,j,k}} \qquad (5)$$

[4] shorturl.at/eiptM

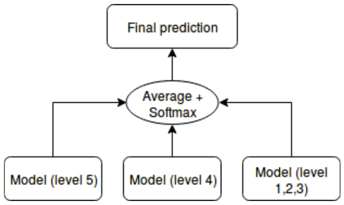


Figure 3: Final model for the prediction of the SMR. Each sub-model is specialized in predicting the level(s) specified in parentheses. Each model is a hybrid architecture.

In each layer, entries with a value greater than $A_{i,j}$ are the most important. Here are some rules extracted from the NN and the decision tree:

- **If** Arousal < 0.228 and Sad < 0.02 **Then** play soft music.
- **If** the player has not visited the NPCs of Level 2 or 5, **Then** force him to do so in the next dilemmas.

Experiments

First, we evaluated the performance of the player's model to accurately predict the level of socio-moral reasoning. We then integrated this model into the game, as well as the adaptation rules. A first experiment was conducted with the non-adaptive version of the game. This experience was carried out on 30 participants aged 8 to 19 years. A second experiment was then conducted with the adaptive version of the game, in which we integrated the learner's model and the adaptation rules. 40 participants aged 10 to 17 participated in the second study. We measured the difference between these two systems in terms of learning gain, immersion, and satisfaction. Before the experiment, participants are asked to pass a pre-test (on a computer) which consisted of 3 dilemmas pre-selected by our team. For each dilemma, the participant listens to the dilemma and then gives his or her opinion verbally. The purpose of the pre-test was to evaluate the SMR level before they start to play, which would allow us to evaluate the impact of the game on the learning. Once the pre-test was completed, the participant could then play the game. Emotions were used in real-time in the adaptive version of the game. Once the game was ended, participants were asked to complete a final questionnaire on immersion, learning, and satisfaction. Once the questionnaire was completed, participants conducted a post-test (similar to the pre-test). We have divided the evaluation of the experimental results into two parts:

- **Subjective Evaluation** : The goal is to evaluate how the player perceives the game in terms of immersion, learning and satisfaction. To do this, we mainly use the answers to the final questionnaires.
- **Objective evaluation**: The goal is to objectively evaluate the learning and satisfaction of the game.

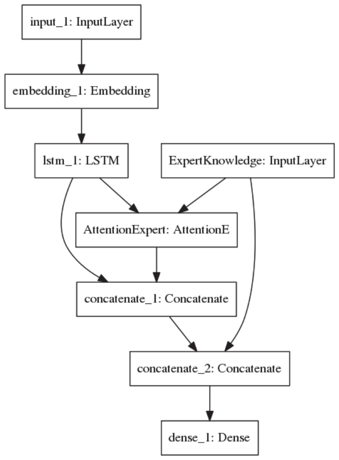

input_1: InputLayer
embedding_1: Embedding
lstm_1: LSTM
ExpertKnowledge: InputLayer
AttentionExpert: AttentionE
concatenate_1: Concatenate
concatenate_2: Concatenate
dense_1: Dense

Figure 4: The proposed hybrid architecture using an LSTM for predicting the SMR

Table 2: Precision, Recall, F1-score and Accuracy of all the trained models for the prediction of the SMR.

Models	Precision	Recall	f1score	Acc
Expert-Only	0.47	0.40	0.38	0.40
cnn-only	0.58	0.53	0.49	0.53
lstm-only	0.42	0.43	0.42	0.43
cnn-expert	0.62	0.62	0.62	0.62
lstm-expert	0.54	0.53	0.51	0.53
cnn-expert-att	0.67	0.65	0.63	0.65
lstm-expert-att	0.59	0.60	0.58	0.60
Final model	**0.72**	**0.75**	**0.73**	**0.75**

Evaluation of the hybrid model

The data consists of 731 verbatims (in French) manually annotated by experts. Since the data set is unbalanced (class 4 and 5 have fewer examples), we have trained different specialized models in the prediction of each of the levels with fewer examples. We then used ensemble methods to combine the results of these models and produce the final prediction. The architecture of the final model is presented in figure 3. Figure 4 shows the architecture we propose for predicting the level of socio-moral reasoning. The models take as input the verbatims that have been pre-processed (tokenization, text to sequence, etc.) and vectorized. The vectors are then sent to the *embbeding* layer. In the figure 4, vectors from the embedding layer are sent to the LSTM layer (note that we only considered the output of the last cell). The a priori knowledge and the output of the LSTM are then passed to the attention layer which performs a combination of the two data sources (as presented above). The vector extracted from the attention layer is then fused with the a priori knowledge and output of the LSTM. The concatenation of these three data sources is passed to the last layer for final prediction. This process is the same for the CNN, except that the attention layer takes as input the a priori knowledge and the result of the *pooling* operation applied at the CNN output. To assess the added value of our solution, we considered two similar models, the *cnn-expert* and the *lstm-expert*. However, these models do not have an attention layer based on a priori knowledge. The data are concatenated with the a

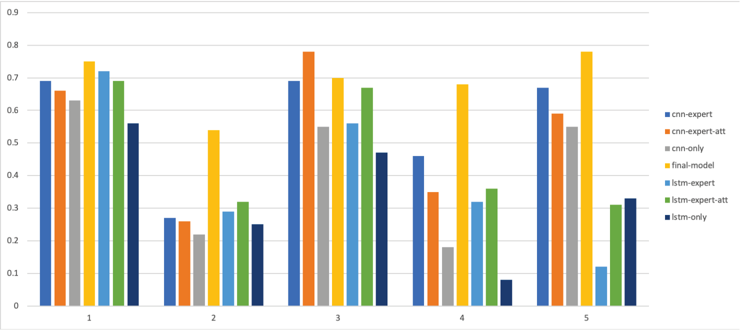
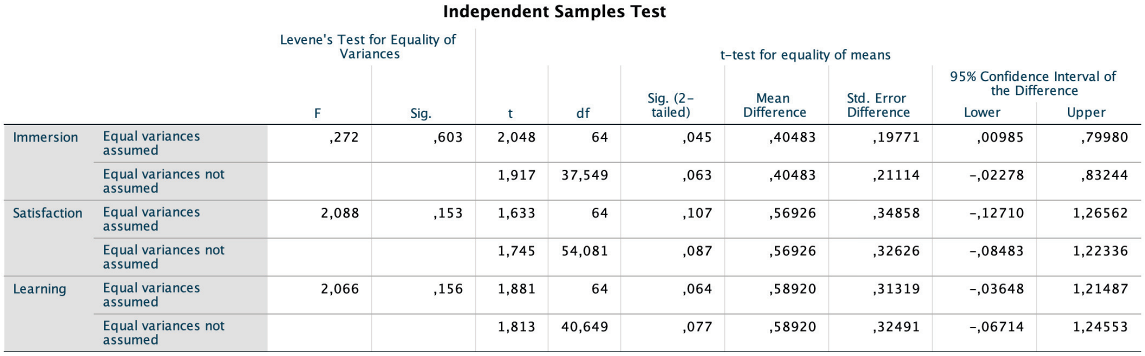


Figure 5: F1-score of all models for each SMR level.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for equality of means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Immersion	Equal variances assumed	,272	,603	2,048	64	,045	,40483	,19771	,00985	,79980
	Equal variances not assumed			1,917	37,549	,063	,40483	,21114	-,02278	,83244
Satisfaction	Equal variances assumed	2,088	,153	1,633	64	,107	,56926	,34858	-,12710	1,26562
	Equal variances not assumed			1,745	54,081	,087	,56926	,32626	-,08483	1,22336
Learning	Equal variances assumed	2,066	,156	1,881	64	,064	,58920	,31319	-,03648	1,21487
	Equal variances not assumed			1,813	40,649	,077	,58920	,32491	-,06714	1,24553

Figure 6: Comparison between the subjective evaluation of the non-adaptive version and the adaptive version of the game, using a *T test* with independent samples.

priori vector.

All models were trained on 80% (including 20% as validation data) of the data and tested on the remaining 20%. The results are presented in the table 2 and figure 5. As we can see, models that take into account a priori knowledge give good results in predicting classes with few samples compared to other models (see f1score for classes 2, 4, and 5). This suggests that integrating expert knowledge into neural models improves classification even when the data set is unbalanced. Overall, models using CNN have worked better than models using the LSTMs since it has been shown that the latter is a more generalizable solution when there is a lot of data available. We have therefore integrated the CNN-based model into the game. The prediction of the SMR level in the game is done in real-time. We first record the learner's audio justification before transforming it into text using Google speech to text API[5]. The text is then sent to the model that makes the prediction and updates the learner's model.

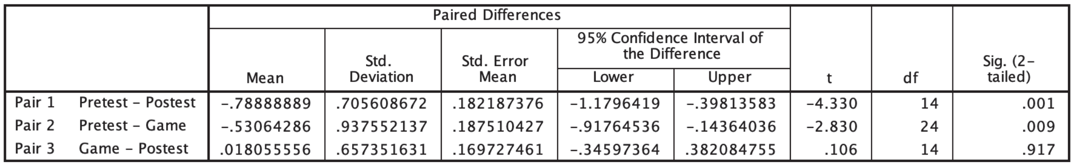
Subjective evaluation

Evaluating learning : The subjective assessment of learning aims to know what the player thinks about his learning in the game. Explicit questions such as: '*did you feel like you learned things in the game?*' were asked. The answers were presented in the form of a Likert scale ranging from 1 to 6 (corresponding respectively to not having learned something and to have learned something). Figure 6 shows the difference in the average learning rate (according to the players) between the two versions of the game. We can see that there is a difference between the two versions even if it is not statistically significant ($p = 0.064$).

Evaluating immersion : The subjective evaluation of immersion aims to know what the player thinks about immersion in the game. Explicit questions such as: '*I was sometimes so involved that I forgot I was in a game*' were asked. We can see in the figure 6 that there is a significant difference ($p = 0.045$) in immersion (according to the players) between the two versions. This means that the adaptive version is significantly more immersive than the non-adaptive version. It should be noted that in the adaptive version, the background music could change according to emotions, which is one of the factors that probably contributed to this result.

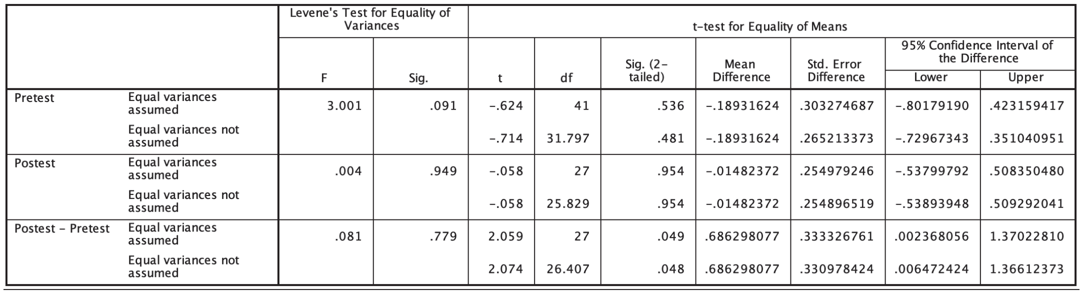
Evaluating satisfaction: The subjective assessment of the satisfaction aims to assess what the player thinks about the game in general. Explicit questions such as: '*Did you like playing this game?* ' were asked. As we can see in figure 6, there is a difference in satisfaction between the two versions. However, this result is not significant since $p = 0.10$. We cannot, therefore, draw any conclusions for this dimension. Perhaps with more participants, we would see the difference. More investigations should be conducted.

[5] https://cloud.google.com/speech-to-text/



		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Pretest - Postest	-.78888889	.705608672	.182187376	-1.1796419	-.39813583	-4.330	14	.001
Pair 2	Pretest - Game	-.53064286	.937552137	.187510427	-.91764536	-.14364036	-2.830	24	.009
Pair 3	Game - Postest	.018055556	.657351631	.169727461	-.34597364	.382084755	.106	14	.917

Figure 7: Assessment of learning in the game : Paired samples test.



		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pretest	Equal variances assumed	3.001	.091	-.624	41	.536	-.18931624	.303274687	-.80179190	.423159417
	Equal variances not assumed			-.714	31.797	.481	-.18931624	.265213373	-.72967343	.351040951
Postest	Equal variances assumed	.004	.949	-.058	27	.954	-.01482372	.254979246	-.53799792	.508350480
	Equal variances not assumed			-.058	25.829	.954	-.01482372	.254896519	-.53893948	.509292041
Postest - Pretest	Equal variances assumed	.081	.779	2.059	27	.049	.686298077	.333326761	.002368056	1.37022810
	Equal variances not assumed			2.074	26.407	.048	.686298077	.330978424	.006472424	1.36612373

Figure 8: Comparison of the 2 versions of the game : Independent samples test.

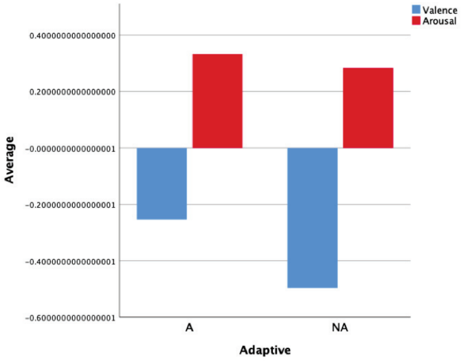


Figure 9: Visualization of the difference (comparison of means) between the valence and arousal (p<0.001) between the non-adaptive (NA) and adaptive (A) versions of the game.

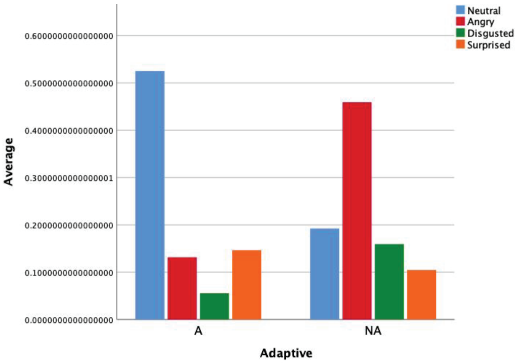


Figure 10: Visualization of the difference (comparison of means) of emotions ($p < 0.001$) between the non-adaptive (NA) and adaptive (A) version of the game.

Objective evaluation

Evaluating learning : To measure the potential of the game in supporting users to develop a higher level of SMR, we will compare the average of the SMR levels obtained by the players during the pre-test, the post-test and the play session. We eliminated participants whose data was partial and/or biased. The results show that the average SMR level of players during the pre-test (1.70) is lower than their average SMR levels during the play session (2.30) and the posttest (2.53) in a very significant way ($p = .001$) (figure 7). Thus, the game allowed a significant increase in the SMR. We also compared the non-adaptive and adaptive versions of the game. In figure 8), the difference between the posttest and the pretest is higher ($p = 0.048$) for those who played the adaptive version compared to those who played the non-adaptive version. Note that, the initial SMR levels are significantly lower for those who played the adaptive version compared to others. This is because participants who played the non-adaptive version were older. Thus, the adaptive version of the game was more effective in supporting the learning of socio-moral reasoning than its non-adaptive version.

Evaluating emotions : Emotions promote learning (Tyng et al. 2017). For all the participants (of the adaptive version), we averaged their emotional reactions to each dilemma. Neutral and anger emotions tend to be more present in activities involving reading. Also, these emotions tend to manifest with greater intensity than others when captured with the Facereader (Alitalo 2016). In LesDilemmes, the most common activity is reading. This is even truer in the adaptive version of the game since we have added textual feedbacks (learning messages, etc). The game does not have a dynamic similar to ”real video games” in the sense that the main character has no liberty in the environment except to make decisions, give his opinion and evaluate others through clicks.

Thus, we will generally observe emotions that are more negatives than positives. We made a first comparison of the means between the valence and the arousal (see figure 9) on the 2 versions of the game. We see that the mean valence is significantly higher in the adaptive (A) version than in the non-adaptive (NA) version. The adaptive version therefore generated ($p < 0.001$) more positive emotions than the non-adaptive version despite the presence of more ”textual” content.

We made a second comparison of the means (see figure 10) involving some of the 7 basic emotions. As shown in that figure, the non-adaptive version generated significantly more anger and disgust than the adaptive version ($p < 0.001$). On the other hand, the adaptive version evoked more neutrality and surprise among the players. Several studies have shown that surprise is an emotion that plays a major role in learning (Foster and Keane 2019). In fact, elements causing surprise are stored more easily and recalled more accurately than elements causing less surprise.

Conclusion

We developed a prototype of a serious game for SMR development. The assessment of the game suggests that it was appreciated by the players in terms of immersion, playability, and learning. The adaptive version of the game can keep the player in an emotional state appropriate for learning. This is possible thanks to the model that automatically predicts the SMR and the adaptation rules. Although there is still room for improvement in the game itself (game dynamics), the adaptive version had a significantly more positive effect than the non-adaptive version on all dimensions assessed, and we are convinced that this positive impact would have been even higher if the automatic transcription from audio to text were more accurate.

Acknowledgments

We would like to thank the NSERC (National Science and Engineering Council) of Canada and the ABCs Developmental Neuropsychology Lab of the University of Montreal.

References

Alitalo, T. 2016. Using facereader to recognize emotions during self-assessment relating to dyslexia.

Andrews, A. 2011. Serious games for psychological health education. In *International Conference on Virtual and Mixed Reality*, 3–10. Springer.

Beauchamp, M.; Dooley, J. J.; and Anderson, V. 2013. A preliminary investigation of moral reasoning and empathy after traumatic brain injury in adolescents. *Brain injury* 27(7-8):896–902.

Bologna, G., and Hayashi, Y. 2018. A comparison study on rule extraction from neural network ensembles, boosted shallow trees, and svms. *Applied Computational Intelligence and Soft Computing* 2018.

Chiasson, V.; Vera-Estay, E.; Lalonde, G.; Dooley, J.; and Beauchamp, M. 2017. Assessing social cognition: age-related changes in moral reasoning in childhood and adolescence. *The Clinical Neuropsychologist* 31(3):515–530.

Chorowski, J., and Zurada, J. M. 2015. Learning understandable neural networks with nonnegative weight constraints. *IEEE transactions on neural networks and learning systems* 26(1):62–69.

Coro, G.; Pagano, P.; and Ellenbroek, A. 2013. Combining simulated expert knowledge with neural networks to produce ecological niche models for latimeria chalumnae. *Ecological modelling* 268:55–63.

Damashek, M. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267(5199):843–848.

Dondlinger, M. J. 2007. Educational video game design: A review of the literature. *Journal of applied educational technology* 4(1):21–31.

Dooley, J. J.; Beauchamp, M.; and Anderson, V. A. 2010. The measurement of sociomoral reasoning in adolescents with traumatic brain injury: A pilot investigation. *Brain Impairment* 11(2):152–161.

Ekman, P. 1999. Basic emotions. *Handbook of cognition and emotion* 98(45-60):16.

Foster, M. I., and Keane, M. T. 2019. The role of surprise in learning: Different surprising outcomes affect memorability differentially. *Topics in cognitive science* 11(1):75–87.

Granic, I.; Lobel, A.; and Engels, R. C. 2014. The benefits of playing video games. *American psychologist* 69(1):66.

Kohlberg, L. 1984. Essays on moral development: The psychology of moral development (vol. 2).

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, 957–966.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Murdoch, W. J., and Szlam, A. 2017. Automatic rule extraction from long short term memory networks. *arXiv preprint arXiv:1702.02540*.

Ryan, R. M., and Deci, E. L. 2008. A self-determination theory approach to psychotherapy: The motivational basis for effective change. *Canadian Psychology/Psychologie canadienne* 49(3):186.

Towell, G. G., and Shavlik, J. W. 1994. Knowledge-based artificial neural networks. *Artificial intelligence* 70(1-2):119–165.

Tyng, C. M.; Amin, H. U.; Saad, M. N.; and Malik, A. S. 2017. The influences of emotion on learning and memory. *Frontiers in psychology* 8:1454.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.