Constrained Multimodal Neural Networks for Interpretable Adaptation Rule Extraction in Game-Based Learning

Ange Tato
Université Laval
Québec, Canada
ange-adrienne.nyamen-tato@fse.ulaval.ca

Roger Nkambou
Université du Québec à Montréal
Montréal, Canada
nkambou.roger@uqam.ca

ABSTRACT

Establishing adaptation rules is a key step in designing effective adaptive systems and is traditionally carried out by domain experts. This paper presents a novel approach based on a constrained multimodal neural network (NN) to extract actionable adaptation rules from learner data. Unlike post-hoc explainability methods, the architecture is intrinsically interpretable through non-negative weight constraints, enabling stable and globally consistent identification of adaptation levers. The extracted patterns are validated using a decision tree and operationalized into concrete rules regulating feedback, NPC reactions, and emotional scaffolding in a serious game. An empirical evaluation comparing adaptive and non-adaptive versions shows improved learning outcomes (with marginal statistical significance, $p = 0.058$) and significantly enhanced emotional states conducive to learning ($p < 0.001$).

Keywords

Adaptation Rules, Socio-moral Reasoning, Neural Networks, Decision Tree, Game-Based Learning

1. INTRODUCTION

Adaptive game-based learning systems personalize learning by adapting to learners' cognitive, behavioral, and affective states [1]. Central to these systems are *adaptation rules* that govern feedback and pedagogical interventions, yet they are typically expert-defined, limiting scalability and responsiveness to multimodal data [14].

Prior work has explored adaptive mechanisms using knowledge tracing, Bayesian models, and deep learning [11, 6], but often focuses on task selection while overlooking affective and interactional dimensions critical for engagement in serious games [15]. In open-ended environments, adaptation must also regulate feedback, emotional engagement, and character interactions [20].

Ange Tato, and Roger Nkambou. Constrained Multimodal Neural Networks for Interpretable Adaptation Rule Extraction in Game-Based Learning. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 622–626. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039675

Neural networks (NNs) can model complex multimodal data but remain difficult to translate into actionable pedagogical rules [8, 4]. Existing rule extraction methods rely on post-hoc explanations or expert-driven design. In contrast, we propose a constrained multimodal NN that embeds interpretability through non-negative weights, enabling stable and globally consistent rule extraction.

This paper makes three contributions: (1) a constrained multimodal NN architecture with intrinsic interpretability; (2) a rule extraction pipeline combining weight analysis and decision tree validation; and (3) an empirical evaluation in a serious game context.

More broadly, we show that constraining neural architectures can transform them into interpretable rule generators, bridging data-driven modeling and actionable pedagogical decision-making.

2. LESDILEMMES: GAME CONTEXT

LesDilemmes [17, 16] is a serious game designed to assess and develop socio-moral reasoning in youths (ages 8–18). Players engage with dilemmas, justify decisions, and evaluate NPC responses spanning five reasoning levels from the So-Moral framework [3]: (1) Authoritarian, (2) Egocentric, (3) Interpersonal, (4) Societal, and (5) Evaluative. Intermediate levels (1.5–4.5) capture developmental transitions. These levels are modeled as developmental constructs independent of age, enabling joint analysis across participants.

The learner model integrates affective state, cognitive profile, and socio-moral reasoning. Game dynamics combine social feedback (e.g., "likes") with pedagogical rules governing NPC reactions and feedback.

Expert-defined rules regulate interactions. For instance, agreement with a higher-level NPC ($r_n > r_p$) triggers positive feedback, while learners below level 4 receive encouragement and scaffolding toward higher reasoning. These rules may co-occur, reflecting authentic pedagogical practice.

3. CONSTRAINED MULTIMODAL NEURAL NETWORK

3.1 Architecture

We use a multilayer perceptron (MLP) as a structured aggregation mechanism for multimodal learner data. All weights are constrained to be non-negative, ensuring intrinsic interpretability [5]. All inputs are normalized to $[0, 1]$.

Given a neuron with inputs x_1, x_2 and non-negative weights w_1, w_2:

$$y = f(w_1 \cdot x_1 + w_2 \cdot x_2 + b) \quad (1)$$

where f is ReLU. Under non-negativity, feature importance is directly reflected by weight magnitude ($w_1 > w_2 \Rightarrow x_1$ is more influential).

The model integrates three modalities (Fig. 1): (1) emotions (9 features), (2) NPC evaluations (5 binary features), and (3) gameplay context (reasoning level of the first NPC visited). The output layer uses Softmax to predict $[1, 0]$ (reasoning ≤ 3) or $[0, 1]$ (reasoning > 3). A dropout layer prevents overfitting given the limited dataset.

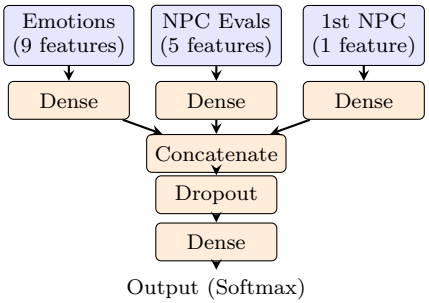

Figure 1: Constrained multimodal NN architecture for rule extraction.

3.2 Feature Importance Computation

To mitigate stochastic variability, 20 training runs are performed. Feature importance $A_{i,j}$ for input i predicting output j is computed using softmax normalization of penultimate-layer weights:

$$a_{i,j} = \frac{e^{w_{i,j}}}{\sum_{i=1}^{n} e^{w_{i,j}}} \quad (2)$$

$$A_{i,j} = \frac{\sum_k a_{i,j,k}}{\sum_k \sum_j a_{i,j,k}} \quad (3)$$

where k indexes training runs. Higher $A_{i,j}$ indicates greater influence on the prediction of output j.

This design transforms the neural network from a predictive black box into a structured aggregation mechanism whose weights directly encode interpretable relationships between learner features and outcomes. As a result, the model supports global rule extraction rather than local post-hoc explanations.

4. RULE EXTRACTION AND VALIDATION

4.1 Neural Network Analysis

Table 1 reports the penultimate-layer weights and averaged importance $A_{i,j}$ across 20 runs. *Emotions* consistently show the highest importance ($A_{i,1} = 0.475$, $A_{i,2} = 0.689$), outweighing NPC evaluations and first NPC visited.

The predictive performance of the model reached an average classification accuracy between 65% and 70% depending on the training run, which is consistent with the exploratory nature of the study and the limited dataset size. Importantly, the NN is not used as a final predictive model but as a mechanism for identifying stable explanatory patterns across modalities—which is why the final adaptation rules combine NN importance rankings with decision tree thresholds. This hybrid strategy compensates for the NN's inability to directly specify threshold values while leveraging its capacity to model nonlinear multimodal interactions.

Table 1: CMNN penultimate layer weights from 3 runs. $A_{i,j}$ averaged over 20 runs.

	Run 1		Run 2		Run 3		$A_{i,j}$	
Modality	[1,0]	[0,1]	[1,0]	[0,1]	[1,0]	[0,1]	[1,0]	[0,1]
Emotions	.007	.812	.000	.751	.925	.739	.475	.689
Eval NPCs	.066	.000	.323	.040	.043	.378	.367	.276
1st NPC visited	.067	.607	.742	.099	.116	.071	.158	.035

Table 2 details importance per emotion. **Arousal** is the most influential ($A_i = 0.234$), followed by sadness (0.165), surprise (0.155), and happiness (0.120). Table 3 shows that NPCs at levels 2 and 5 are most predictive ($A_i = 0.215$ and 0.273 respectively).

Table 2: Emotion branch weights and importance A_i (20 runs).

Emotion	Run 1	Run 2	Run 3	A_i
Neutral	.080	.093	.000	.001
Happy	.000	.475	.000	.120
Sad	.672	.013	.333	.165
Angry	.069	.702	.000	.090
Surprised	.000	.034	.583	.155
Scared	.138	.000	.041	.001
Disgusted	.000	.016	.001	.106
Valence	.346	.006	.045	.128
Arousal	.039	.404	.578	**.234**

Table 3: NPC evaluation weights and importance A_i (20 runs).

NPC Level	Run 1	Run 2	Run 3	A_i
Eval 1	.939	.186	.000	.198
Eval 2	.078	.755	.000	**.215**
Eval 3	.052	.000	.938	.167
Eval 4	.028	.588	.000	.147
Eval 5	.920	.369	.000	**.273**

4.2 Decision Tree Validation

A decision tree (Fig. 2) trained on the same dataset validates the NN findings and provides threshold values for operationalization. Key extracted rules:

- Low *arousal* $\Rightarrow$ higher probability of reasoning level > 3
- Agreement with NPC level 2 + low disgust $\Rightarrow$ reasoning level > 3
- No agreement with NPC level 2 + low surprise $\Rightarrow$ reasoning level < 3
- No agreement with NPC level 2 + agreement with NPC level 3 $\Rightarrow$ reasoning level > 3

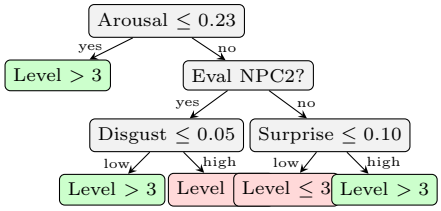

Figure 2: Simplified decision tree for socio-moral reasoning prediction.

These findings motivate concrete game adaptations: background music is changed according to arousal and valence, and players are required to visit all NPCs (including levels 2 and 5) rather than stopping after three.

While the decision tree is trained on the same dataset, it provides an independent, symbolic approximation of the learned relationships, enabling the identification of explicit thresholds and facilitating operational deployment. One advantage of using the NN is that, as the data grows, the model will still be able to highlight features that need to be observed and modified to improve the mastery of the knowledge, compared to the decision tree alone.

5. EMPIRICAL EVALUATION

5.1 Dataset and Protocol

29 youths (ages 8–18) played the non-adaptive version of *LesDilemmes* across 9 dilemmas, yielding 261 observations. Verbal justifications were transcribed via Google Speech-to-Text and analyzed with a data mining algorithm. Emotional data were collected via *Facereader* (7 basic emotions + valence + arousal). An adaptive version incorporating the extracted rules was subsequently developed and evaluated against the baseline.

5.2 Learning Outcomes

Table 4 reports paired t-test results. Players showed significant improvement from pre-test (mean 1.70) to in-game performance (2.30, $p = 0.029$) and post-test (2.49, $p < 0.001$). The non-significant difference between in-game and post-test ($p = 0.445$) indicates that gains persist after gameplay.

Table 4: Paired-samples t-test results for learning in *LesDilemmes*.

Comparison	Mean diff.	SD	t	N	p
Pre → Post	−0.789	0.706	−4.33	14	< .001
Game → Post	−0.233	1.188	−0.78	15	.445
Pre → Game	−0.491	1.056	−2.32	24	.029

Comparing adaptive (A) vs. non-adaptive (NA) versions (Table 5), the pre–post gain was 0.789 for A vs. 0.333 for NA ($p = 0.058$). Note that A participants had a lower pre-test baseline (1.70 vs. 2.18), which is partly explained by the NA group being older on average.

5.3 Emotional Impact of Adaptation

Fig. 3 compares emotional profiles between conditions. The adaptive version yields significantly higher valence ($p < 0.001$), confirming that adaptation rules oriented toward positive affect were effective despite the reading-heavy nature of the

Table 5: Adaptive (A) vs. non-adaptive (NA) comparison.

Measure	Version	N	Mean	SD
Pre→Post gain	A	15	0.789	0.706
	NA	29	0.333	0.749
Pre-test	A	15	1.700	0.575
	NA	29	2.178	0.675
Post-test	A	16	2.536	0.684
	NA	29	2.511	0.648

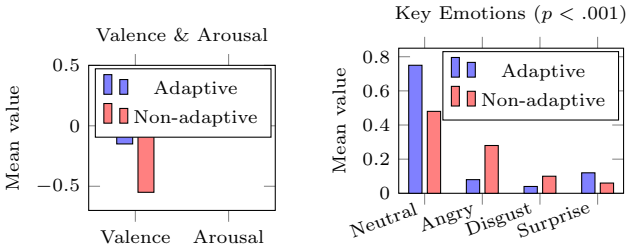

Figure 3: Emotional comparison between adaptive (A) and non-adaptive (NA) versions ($p < 0.001$). The adaptive version produces significantly higher valence, lower anger and disgust, and higher surprise.

game. The emotions neutral and anger are typically more prevalent during reading activities and tend to manifest with greater intensity when measured via Facereader [18]. In *LesDilemmes*, the predominant activity is reading—especially true in the adaptive version with additional learning and encouragement messages—and the main character has no direct control over the environment, only making decisions and evaluating others through button clicks. Therefore, negative emotions being more frequent than positive ones is entirely expected. Anger and disgust were significantly reduced in the adaptive condition, while neutrality and surprise increased. Surprise is particularly relevant: research shows surprising stimuli are remembered faster and recalled more accurately [7], and it promotes deeper cognitive engagement [2].

6. ADAPTATION RULES

A central contribution of this work is the translation of NN-derived patterns and decision tree thresholds into concrete, actionable rules deployed in *LesDilemmes*. Table 6 summarizes the five main rules implemented in the adaptive version, along with their source (NN or DT), the targeted learner state, and the game action triggered.

These rules complement expert-defined rules by targeting emotional regulation as a primary lever—an aspect absent from the original design. The NN-derived rules are globally stable (extracted from 20 runs), making them robust to stochastic training variability and suitable for deployment without per-learner retraining.

7. DISCUSSION

A key insight of this work is that architectural constraints can play a central role in making neural models usable in educational settings. Rather than relying on post-hoc interpretability, constraining the hypothesis space enables the

Table 6: Operationalized adaptation rules in *LesDilemmes*.

Condition	Source	Game Action	Goal
Valence < 0	NN (arousal)	Switch to cheerful music	Mood lift
Valence ≥ 0	NN (valence)	Play soft background music	Maintain state
Agrees with NPC $r_n > r_p$	Expert rule	NPC shows happy face + thumbs-up	Reinforce
$r_p < 4$	Expert rule	Display encouraging + scaffolding message	Scaffold
Player skips NPCs 2 or 5	NN (Eval 2/5)	Force exploration of remaining NPCs	Ensure exposure

extraction of stable and pedagogically meaningful rules that can be directly operationalized.

7.1 Interpretability by Design

A key methodological contribution of this work is the distinction between *intrinsic* and *post-hoc* interpretability. Techniques such as LIME [12] and SHAP [10] explain individual predictions after training, providing local approximations that may vary across instances. In contrast, non-negative weight constraints make the trained model itself interpretable at a global level: because all weights $w_i \geq 0$ and inputs $x_i \in [0, 1]$, the relative magnitude of weights directly encodes feature importance without auxiliary approximation.

This property is especially valuable in educational contexts, where adaptation rules must be stable, auditable, and pedagogically justifiable. A rule derived from a globally consistent model carries more credibility for instructional designers than a saliency map computed locally for a single learner episode.

7.2 Role of Emotions in Socio-Moral Learning

The prominence of affective features—particularly arousal, surprise, and happiness—in predicting reasoning level has important theoretical implications. It aligns with the cognitive-affective learning model [19], which posits that emotional states modulate attention, working memory, and long-term retention. High arousal without positive valence can hinder deliberate reasoning, while moderate arousal paired with positive affect creates an optimal state for engagement and higher-order thinking. Positive emotions also increase learner engagement, facilitate learning, and contribute to better long-term retention [13].

The observed increase in surprise in the adaptive condition is consistent with theories of *productive confusion* [7, 2]: unexpected feedback or NPC responses prompt learners to re-evaluate their assumptions, fostering deeper moral reflection. Designing NPCs to occasionally challenge learners with unexpected high-level reasoning—rather than always providing confirmatory feedback—may therefore be a principled lever for supporting reasoning progression.

7.3 Comparison with Related Approaches

Unlike deep knowledge tracing [11], which models latent skill states through LSTM networks but provides little interpretability, our constrained MLP explicitly surfaces which input dimensions drive predictions. Compared to Bayesian student models [6], which require strong prior assumptions, our approach learns directly from multimodal data without structural priors, while still producing transparent outputs. The hybrid NN–decision tree pipeline combines the representational power of neural learning with the symbolic clarity of tree-based rules—a complementary strategy supported by recent work on rule extraction from neural networks [8]. Unlike narrative personalization [20] and goal recognition [9] approaches that require manually authored rules, our rules emerge directly from data via the constrained architecture, making the approach scalable to new game contexts.

7.4 Limitations and Future Directions

Several limitations constrain the generalizability of the current results. First, the sample size ($N = 29$, yielding 261 observations) is small. Results should therefore be interpreted cautiously given the limited sample size and group differences, and viewed as preliminary evidence of the effectiveness of the extracted adaptation rules. Furthermore, the marginal significance of the adaptive gain ($p = 0.058$) warrants cautious interpretation. A pre-registered, adequately powered study is needed to confirm the learning benefit of adaptation. Second, the reasoning level labels depend on automatic speech transcription (Google Speech-to-Text) followed by algorithmic coding, introducing noise that could affect both model training and outcome measurement. Third, the age heterogeneity of participants (8–18 years) introduces developmental variance that may interact with adaptation mechanisms in ways not yet captured by the model.

Future work will address these limitations through: (1) larger and more homogeneous samples stratified by age group; (2) improved multimodal fusion incorporating gaze and physiological signals alongside facial expressions; (3) online model updating to enable within-session personalization rather than rule extraction from a fixed prior dataset; and (4) extension of the framework to other serious game contexts beyond socio-moral reasoning, such as scientific argumentation and collaborative problem-solving.

7.5 Broader Implications for Adaptive ITS

A key insight is that architectural constraints shift the neural model from a predictive black box to a knowledge elicitation mechanism, yielding stable, operationalizable rules without post-hoc interpretability. Three implications follow for adaptive ITS. First, interpretability and predictive power need not be traded off: 65–70% accuracy is maintained alongside globally consistent importance rankings. Second, the modular pipeline—NN for *which* features matter, decision tree for *when* and *at what thresholds*—allows independent updates as data or targets change. Third, globally extracted rules (20 averaged runs) are robust to stochastic variability, enabling real-time deployment without per-session retraining. The approach generalizes to any multimodal dataset with a learner outcome variable: architecture

choices adapt to available sensors while the non-negativity constraint preserves interpretability by construction.

8. CONCLUSION

This paper introduced a data-driven pipeline for extracting actionable adaptation rules from a constrained multimodal neural network in a game-based learning context. By enforcing non-negative weight constraints, the architecture transforms the NN into a globally interpretable aggregation mechanism whose weights directly encode feature importance across emotional, interactional, and gameplay modalities—without relying on post-hoc approximations.

Applied to *LesDilemmes*, the approach identifies arousal, surprise, and happiness as dominant predictors of higher socio-moral reasoning, and interactions with NPCs at levels 2 and 5 as the most informative social levers. These findings, validated by a decision tree analysis, were operationalized into five concrete adaptation rules governing background music, NPC feedback, and player navigation.

Empirical evaluation demonstrates that the resulting adaptive version shows larger pre–post learning gains with marginal statistical significance ($p = 0.058$; adaptive gain 0.789 vs. 0.333 for non-adaptive), substantially improves emotional climate ($p < 0.001$), and generates higher levels of surprise—an emotion associated with deeper cognitive engagement and better retention.

Beyond the specific game context, this work demonstrates that constraining neural architectures can transform them into interpretable rule extraction mechanisms, bridging the gap between data-driven learning and pedagogical design. This positions constrained multimodal neural networks as a practical and principled foundation for adaptive intelligent tutoring systems.

9. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC).

10. REFERENCES

[1] S. Adipat, K. Laksana, K. Busayanon, A. Asawasowan, and B. Adipat. Engaging students in the learning process with game-based learning: The fundamental concepts. *International Journal of Technology in Education*, 4(3):542–552, 2021.

[2] J. E. Adler. Surprise. *Educational Theory*, 58(2):149–173, 2008.

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

[4] G. Bologna and Y. Hayashi. A comparison study on rule extraction from neural network ensembles, boosted shallow trees, and svms. *Applied computational intelligence and soft computing*, 2018(1):4084850, 2018.

[5] J. Chorowski and J. M. Zurada. Learning understandable neural networks with nonnegative weight constraints. *IEEE transactions on neural networks and learning systems*, 26(1):62–69, 2014.

[6] C. Conati. Bayesian student modeling. In *Advances in intelligent tutoring systems*, pages 281–299. Springer, 2010.

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

[8] C. He, M. Ma, and P. Wang. Extract interpretability-accuracy balanced rules from artificial neural networks: A review. *Neurocomputing*, 387:346–358, 2020.

[9] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *Ai Magazine*, 34(3):27–41, 2013.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[11] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[13] L. Shen, M. Wang, and R. Shen. Affective e-learning: using" emotional" data to improve learning in pervasive learning environment. *J. Educ. Technol. Soc.*, 12(2):176–189, 2009.

[14] V. Shute and B. Towle. Adaptive e-learning. In *Aptitude*, pages 105–114. Routledge, 2018.

[15] L. Sun, M. Kangas, and H. Ruokamo. Game-based features in intelligent game-based learning environments: A systematic literature review. *Interactive Learning Environments*, 32(7):3431–3447, 2024.

[16] A. Tato and R. Nkambou. Infusing expert knowledge into a deep neural network using attention mechanism for personalized learning environments. *Frontiers in Artificial Intelligence*, 5:921476, 2022.

[17] A. Tato, R. Nkambou, A. Dufresne, and M. H. Beauchamp. Convolutional neural network for automatic detection of sociomoral reasoning level. *International Educational Data Mining Society*, 2017.

[18] V. Terzis, C. N. Moridis, and A. A. Economides. Measuring instant emotions during a self-assessment test: the use of facereader. In *Proceedings of the 7th international conference on methods and techniques in behavioral research*, pages 1–4, 2010.

[19] C. M. Tyng, H. U. Amin, M. N. Saad, and A. S. Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:235933, 2017.

[20] P. Wang, J. P. Rowe, W. Min, B. W. Mott, and J. C. Lester. Interactive narrative personalization with deep reinforcement learning. In *IJCAI*, pages 3852–3858, 2017.