# Towards General Models of Player Affect

Elizabeth Camilleri
Institute of Digital Games
University of Malta
elizabeth.camilleri.12@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games
University of Malta
georgios.yannakakis@um.edu.mt

Antonios Liapis
Institute of Digital Games
University of Malta
antonios.liapis@um.edu.mt

***Abstract*—While the primary focus of affective computing has been on constructing efficient and reliable models of affect, the vast majority of such models are limited to a *specific* task and domain. This paper, instead, investigates how computational models of affect can be *general* across dissimilar tasks; in particular, in modeling the experience of playing very different video games. We use three dissimilar games whose players annotated their arousal levels on video recordings of their own playthroughs. We construct models mapping ranks of arousal to skin conductance and gameplay logs via preference learning and we use a form of cross-game validation to test the generality of the obtained models on unseen games. Our initial results comparing between *absolute* and *relative* measures of the arousal annotation values indicate that we can obtain more general models of player affect if we process the model output in an ordinal fashion.**

***Keywords*—general affect modeling; emotion annotation; preference learning; relative annotation; games**

## I. Introduction

Research on general artificial intelligence (AI) has focused primarily on computational systems that are able to perform well on a range of objectively-defined cognitive tasks [1], [2]. Games offer complex yet well-defined problems for exploring the capacities of general AI and, as a result, playing games has been the dominant task for testing general AI capacities over the years. Recent studies, however, have argued that testing AI only through its gameplaying abilities is a very narrow perspective to general intelligence [3], [4]. Evidence from neuroscience further supports this stance, suggesting that emotion is a key *facilitator* of general intelligence [5] and that our affective abilities are not only admissible but necessary factors of our general intelligence [6].

Even though the generality of affective interaction appears to be critical for realizing general AI, the general capacity of affect models has not been a research focus within affective computing (AC) yet. The majority of AC studies focus on the design of reliable and effective models which are tied to a *specific* task within a *specific* domain. In this paper, instead, we take the first steps towards the design of affect models that are more *general*. While we focus particularly on the domain of video games as one of the most promising domains for the realization of the *affective loop* [7], we investigate how an affect model can be general across various different tasks (in this case, modeling the experience of playing various games).

We argue that a model of affect can be general only if both its input and output representations are expressed in general terms; in this paper we focus on the *output* of such a potential general model. Further, we are grounded on the evidenced advantages of ordinal annotation [8]–[10] and ordinal modeling of affect [11], [12] in yielding reliable approximates of the underlying ground truth. Based on these studies, we argue that the ordinal processing of affect annotations can provide us with *relative* expressions of the ground truth which we view, in turn, as an essential step towards achieving accurate general models of affect. Our hypothesis is therefore that processing affect annotations (i.e., the model output) in a relative (i.e., ordinal) fashion yields more general affect models than when processing them in an absolute fashion.

To test our hypothesis we construct computational models which output a measure of the player's arousal based on a player's gameplay metrics and skin conductance. The model is tested for its capacity to predict the arousal of a player across three dissimilar games that vary fundamentally in terms of game genre, game rules, and overall play experience. Selecting games that are so different aims to challenge our models' general modeling abilities and offer a reliable benchmark for testing general affect intelligence. The models are trained via preference learning (using neuroevolution and support vector machines) on continuous arousal annotations provided by the players while watching their video-recorded playthroughs. The annotations are converted into ordinal data by comparing pairs of annotated time windows. We explore two different metrics for comparing between time windows and transforming the continuous arousal annotations into ordinal data: a) the *absolute* metric of mean arousal value and b) the *relative* metric of the arousal's average gradient (average change). The results validate our hypothesis as models trained on the relative arousal data were the only ones that manage to surpass the baseline performance (significantly in some cases), whereas models built on the absolute mean arousal value, at best, only reach the baseline. Based on these initial findings we envisage the application of our approach to building more general affect models across tasks in other domains besides games.

## II. General Affect Modeling in Games & Beyond

Modeling affect has received much attention; this section focuses on the task of general affect modeling within the games domain and beyond with an emphasis on the processing of input (Section II-A) and output (Section II-B).

*Player experience modeling* (PEM) [13] is a rather active area of study dedicated to building computational models that capture how players react to certain games, or the elements and

978-1-5386-0563-9/17/$31.00 

events within them. Although PEM covers a broad range of player experience states, these are typically closely linked with affect. Several studies have attempted to acquire models which predict such states as 'fun', 'challenge,' and 'frustration' (e.g., [14], [15] among many). Although research in modeling player experience grows in volume of studies and interest, most of the aforementioned studies are bound to a *specific* game or game genre that the models were trained and tested on [3], [4]. A small number of researchers have begun pushing towards employing a more general perspective in affect modeling [3], [4] and a small number of studies have already made promising steps towards building general models [16]–[18].

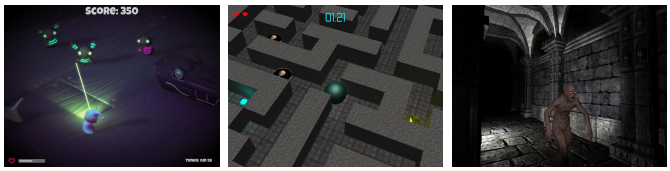
### A. General Input

An affective model able to capture player experience across games should have general features as its input. Such general input can be completely game independent, such as physiological signals. In [16], results showed that average and minimum heart rate as well as 1-step and 2-step differences of skin conductance are good general predictors of self-reported player affect across two different games: a racing game and a 3D maze ball puzzle game. This paper uses a variation of the maze ball puzzle game of [16] for capturing annotations, and also the 1-step differences of skin conductance as input due to this metric's evidenced ability to capture aspects of affect. In a similar study [17], general gameplay and context-based features were devised as input, training a model based on game data of a platformer game and testing the model on a first person shooter game (and vice-versa). General gameplay features are also captured in this study, and are shown to be good predictors for cross-game validation—i.e., testing on an unseen game, similarly to [17]. The varying results of [17] demonstrate how challenging general affect modeling across games can be, which is also corroborated in our findings. Finally, a recent study [18] compared manually-designed features against transfer learning on the same two games of [17], obtaining comparable results as in [17].

### B. General Output

While the input of any machine learning algorithm naturally affects its predictive accuracy, the format of the output (the ground truth of affect) is equally important, if not more. How to annotate affect is not straightforward and approaches vary throughout literature. The predominant approach in affective computing is continuous annotation [8], considering affect on a continuous scale. Affect can be annotated continuously with respect to the affect itself (e.g., via Russell's two-dimensional arousal-valence circumplex model of affect [19]), with respect to time (e.g., via the FeelTrace tool [20]), or both.

Continuous annotation provides richer data in terms of both quality and quantity than discrete annotation [8] but user fatigue and inter-rater disagreements across continuous ratings still prove to be a challenge. In order to derive a *general output* for affective modeling, *relative* and rank-based annotations offer a way to overcome such challenges. Research has shown that people are able to report emotions better in relative terms than in absolute terms [10], [21], and a pairwise rank approach has been used in many studies on affective modeling in games (e.g., [14], [15]). Models in this paper similarly learn the pairwise ranks of affect via preference learning. A real-time discrete rank-based annotation tool has been shown to outperform continuous ratings in terms of inter-rater agreement, while converting real-time continuous rating annotations into ranks was also shown to increase inter-rater agreement [8]. Finally, approaches that consider relative agreements between real-time continuous annotations based on the direction of *change* has proved advantageous over approaches that consider agreement in terms of absolute values, both in terms of inter-rater agreement [10], [22] and in terms of constructing accurate models of affect [23]. For these reasons, in this paper we test and compare an *absolute* and a *relative* approach for processing the annotation data which, in turn, can be used to test the general predictive capacity of our models: the first being the mean value of the arousal trace and the latter being the average gradient value of the arousal trace.




(a) Survival Shooter (SS) (b) Space Maze (SM) (c) Sonancia

Fig. 1. Screenshots from the three different games used in this study.

## III. Testbed Games

Three different games developed in *Unity 3D* (see Fig. 1) were used to test the generality of our affective models: **Survival Shooter** (SS) is a game included in a Unity3D tutorial package.[1] In this top-down game, the player must shoot down as many zombie toys as possible and avoid running out of health due to zombies colliding with him. In the adapted version used in our experiments, the game ends after 60 sec. **Space Maze** (SM) is a maze-based puzzle game [24] where the player, controlling a ball, must collect three diamonds, avoid enemies, and reach a final point before running out of health (by colliding with enemies) or time (90 seconds). In the adapted version used in our experiments, we omit for the sake of simplicity a game feature from the original game whereby the camera changes perspective upon diamond collection. **Sonancia** (Son) is a first-person horror game introduced in [25] where the player must traverse rooms until a room with the objective is reached, while avoiding monsters which can chase and kill the player. Each room has a distinct soundtrack matching a desired level of tension for that room.

Since our study tests how models can generalize across dissimilar games, the three games were chosen due to their differences in genre, game mechanics, camera perspective, graphics style, and the player's goal (or end condition). With

[1] https://www.assetstore.unity3d.com/en/\#!/content/40756

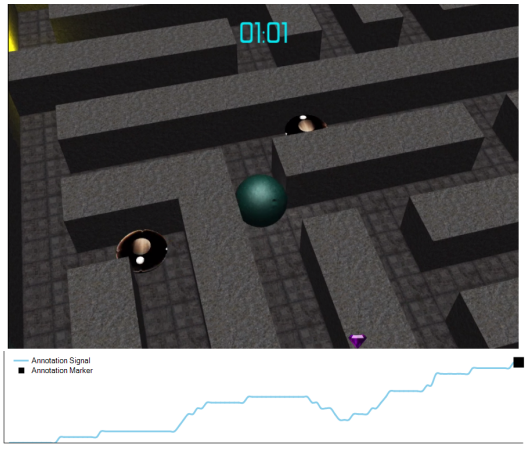



Fig. 2. Screenshot from the *RankTrace* [26] annotation software, showing the recorded playthrough (top) while the entire annotation trace is visible to the user (bottom) who can control the current arousal value (square).

regards to genre, SS is a shooter relying on fast-paced reactions, accurate aiming and constant movement, SM is a physics puzzle that needs accurate timing of movement, and *Sonancia* is a horror game which elicits negative emotions, disorientation and jump scares. Moreover, the camera perspective is top-down in SS, third-person in SM, and first-person in *Sonancia*. The dissimilarities between the three games make them an ideal dataset for testing the capacity of a model to yield general affective models across all games. The game situations, the corresponding gameplay experience and affective responses are so different among the games that a model might not be able to recognize them, perceiving them as unknown or as mere noise.

## IV. Experimental Protocol

Data from all three games was collected using the same protocol. The SS and SM data collection experiments were run by the same participants whereas the *Sonancia* data collection occurred on a different day with other subjects. The data collection process took place over two different settings and at various times throughout the day. However, much care was taken to ensure the same test conditions.

Participants were first briefed about the experiment and then filled in a simple demographic questionnaire. An *Empatica* E4 wristband sensor[2] was then fitted on the player's left wrist to log physiological signals. The subjects were then asked to play each of the SS and SM games twice; the order of which game they played first was randomized to avoid any order biases. In experiments with *Sonancia*, subjects played the same level three times. Before the first playthrough of each game, the player was shown a black screen with game instructions for 30 seconds, during which a baseline for the physiological signals was recorded. During play, game metrics were logged together with a screen-capture video of the game playthrough using the *RankTrace* annotation tool [26]. After each individual playthrough, participants viewed the screen-capture video of their last playthrough and continuously annotated the level of arousal (intensity) characterizing the emotions they recall feeling while playing on an unbounded scale using a USB wheel controller (PowerMate, Griffin Technology) to indicate changes (increase or decrease) in arousal.[3] The *RankTrace* tool [26][4] can visualize and process a plot of the arousal signal being annotated with the USB wheel controller in real-time while the playthrough video is displayed (see Fig. 2).

Only the arousal dimension of affect was annotated in this paper. Obtaining valence annotations as well would have considerably increased the experiment duration and could thus have negatively impacted the quantity and quality of data obtained given the available time and resources. It is also worth noting that the first playthrough of each game was treated as a tutorial for the players to get used to the controls of the game and was therefore not considered in any processing stages described from this point onwards.

## V. Data Collection

A total of 25 participants (10 females) provided data for both SS and SM while 14 participants (5 females) provided data for *Sonancia*. After we applied all data preprocessing steps described in this section, 10, 10 and 11 participant sessions were considered for SS, SM and *Sonancia*, respectively. Participants for SS and SM were aged from 19 to 54 (median age 24) and most of them considered themselves good or expert players of games (70%) while the rest considered themselves novice or non-gamers. As for *Sonancia*, most participants were between 25-34 years old and 36% of participants played games everyday while 45% played frequently or casually; the rest rarely or never played.

For each game session we record and analyze three types of data: annotation traces of arousal, gameplay logs and electrodermal activity. For each data type we follow different preprocessing and feature extraction methods which are detailed in the corresponding sections below.

### A. Annotation Traces

Continuous annotations of arousal were processed on a playthrough-by-playthrough basis. As a preprocessing step, we discarded traces where annotated arousal values were constant. Each playthrough annotation trace was then processed via a sliding window approach, splitting the session into equally-sized time-windows. Several parameters were considered: the *window size* in seconds ($w$), the *step size* ($s$) as the delay between the start of the current window and the start of the previous window, and the *reaction lag* ($l$) as the time in seconds by which the annotation was offset with respect to the corresponding game log and physiological signals to account for a lag in the annotator's response to the events in the playthrough video. We set default values to $w = 3$, $s = 3$ (i.e., no overlap between consecutive windows), and $l = 0$

[2]https://www.empatica.com/e4-wristband

[3]Due to the nature of the horror genre players of *Sonancia* were asked to annotate tension which is often used interchangeably with arousal [19].

[4]Available at: http://www.autogamedesign.eu/software

(i.e., no reaction lag). This resulted in a total of 139 windows for SS, 227 for SM, and 256 for *Sonancia* across all sessions.

Two metrics were calculated for each window: the mean arousal value $\mu_A$ (*absolute* metric) and the average gradient $\Delta_A^1$ (*relative* metric) of the arousal, i.e., the average first differences within the continuous annotation window. Outliers for each of these metrics were handled by capping values outside $\pm 3$ standard deviations to that value. Further, the values for each metric were min-max normalized to $[0, 1]$.

In each playthrough, the annotation windows were used to derive rankings between arousal values within adjacent windows. We assume that annotators have a limited memory and are therefore able to compare only their current arousal to their perceived arousal a few seconds before. The same windows used for annotation ($w$=3, $s$=3, $l$=0) are applied to game logs and EDA signals (described below) to derive rankings of those metrics between adjacent windows.

### B. Gameplay Logs

Based on general game design properties, we designed a number of features which we deemed to be general across multiple types of games. These include: goal-oriented events $G^+$ (i.e., events which lead the player towards the goal), goal-opposed events $G^-$ (i.e., events which lead the player away from the goal), the player's distance traveled $D$ and time spent moving $M$, number of enemies engaged with the player $E$, time since the start of the game $t_S$, and level of player fatigue $t_F$ (i.e., how long the player has been playing in total, including tutorials). Goal-related features $G^+$ and $G^-$ were treated as binary, indicating whether at least one goal-oriented or goal-opposing event occurred in the given window respectively. All other features were treated as scalars.

Unlike the other gameplay features, $G^+$, $G^-$, and $E$ are derived differently for each of the three games. Goal-oriented events (contributing to $G^+$) included hitting and killing enemies for SS, and collecting diamonds for SM. Goal-opposed events (contributing to $G^-$) for both SS and SM consisted of collisions with enemies; $E$ in both SS and SM was the number of enemies visible on screen. For *Sonancia*, $G^+$ events were when a player entered a new room, when monsters lost sight or ceased to chase the player; $G^-$ events were when the monsters hit the player, when the player died, and when monsters gained sight or started to chase the player. Finally, $E$ for *Sonancia* was the number of monsters chasing the player.

It is important to note that gameplay features offer the necessary *context* for reliably inferring the affective state of the player. At the same time, however, they are designed to be as general as possible across games. All gameplay features were normalized with respect to the respective game only using Z-score normalization.

### C. Electrodermal Activity

From the various physiological signals recorded by E4, we only consider the electrodermal activity (EDA)[5] in this study. EDA is measured in micro Siemens ($\mu S$). Considering only sessions with an EDA signal which was not noisy, we extracted four descriptive statistics from the EDA signal for each given window: mean, median, standard deviation, and variance. We then divided each of these features for each window by the corresponding features calculated over the baseline EDA signal. This final feature value represents the *relative* change of the statistic with respect to a particular participant's baseline. Beyond these standard statistical features, and inspired by the study of Holmgård et al. [27], we applied a continuous decomposition analysis (CDA) [28] to the EDA signal using Ledalab. The outcome of CDA is the decomposition of the EDA in its phasic and tonic components. From these components we considered the mean and the integral of the phasic driver, and the mean tonic driver of the signal in each window.

[5]The terms skin conductance and electrodermal activity are used interchangeably in this paper.

## VI. Preference Learning

This work explores how general models of affect can be constructed across dissimilar activities and tasks. We thus use machine learning to discover the mapping between the annotation traces (model output) and the gameplay and EDA (model inputs) features. As the data in annotation metrics is ordinal, we use preference learning [29], [30] to construct our computational models of arousal across the three games.

Preference learning is a machine learning process by which the assumed global order underlying a set of preference ranks is inferred [29]. The output of a preference learning algorithm is a computational model which maps a set of input features that characterize the input instances to the inferred global order [30]. In this paper we use RankSVM [31] and neuroevolutionary [30] preference learning, applying sequential forward feature selection to both. For all experiments reported in this paper we use the Preference Learning Toolbox [32].

The RankSVM preference learning algorithm [31] is used due to its deterministic nature, low computational effort, and capability to reach high levels of performance. RankSVM is an ordinal version of the original Support Vector Machine (SVM) that maps instances to a high-dimensional space and finds a hyperplane which best splits the data into two groups. In this paper RankSVM uses a radial basis function (RBF) kernel with $\gamma = 1$. For comparative purposes we also employ the neuroevolutionary (NE) preference learning algorithm, which uses artificial evolution to adjust the connection weights of a neural network which predicts the ordinal output. It does so by employing a fitness function that rewards matching of preferences. The algorithm has been used extensively in the literature (e.g., see [30] among many). All experiments in this paper used a population size of 200, a uniform crossover probability of 0.8 and a mutation probability of 0.1. For selection, we used a Roulette wheel scheme with 40 parents. Finally, we used an elitism strategy of size 20 and iterated over 50 generations.

### A. Feature Selection

To select the most appropriate input features for our models we use sequential forward selection (SFS). SFS is a hill

climber that starts with an empty set of features and iteratively adds one feature at a time, by trying all features in combination with any already selected features, picking the best combination based on model performance. The process runs until the addition of another feature results in loss of model performance.

## VII. Results and Analysis

In this section we investigate the impact of two different ways of processing the arousal annotation data (ground truth) on the resulting model's capacity to generalize across the three game test-beds. On one hand, we calculate annotation values based on the *absolute* measure of mean annotations ($\mu_A$) within a window. On the other hand, we derive the average gradient ($\Delta_A^1$) of the annotation within each window which is a *relative* metric based on first differences of the annotation data. We use these values as the outputs of a preference learning mechanism that attempts to predict which of two adjacent time windows would have a higher annotation metric. A successful model would be able to predict whether the arousal level in the next time window will increase or decrease and by how much. A total of 46 pairwise ranks for SS, 96 for SM, and 77 for *Sonancia* were obtained with $\mu_A$ while 76 for SS, 152 for SM, and 157 for *Sonancia* were obtained with $\Delta_A^1$. In all experiments reported below, the *baseline* performances are derived by finding which window (at time $t$ or at $t-1$) is preferred more often in all annotations within the test set. The baseline of the dataset is the highest of these two numbers, expressed as a percentage over the sum of the two numbers. Significant differences for both game-specific and cross-game validation accuracies are assessed via their 95% confidence interval bounds.

### A. Game-specific Affective Models

To first validate the algorithms' ability to learn models of affect on the games chosen, both EDA and gameplay features were used as input to learn their mapping with self-reported arousal on a game-by-game basis. While this experiment does not advance the vision of a general model of affect, it nevertheless tests whether there is a game-specific mapping between the features chosen and arousal. If such a highly accurate predictive model exists for one game, perhaps it can be generalizable across two or three games. Indicatively, using RankSVM with $\Delta_A^1$, the obtained leave-one-out cross-validation performances for SS, SM and *Sonancia* are $86.84\%$, $69.08\%$, and $75.16\%$, respectively; all values are significantly above their corresponding baselines of $53.95\%$, $51.97\%$ and $51.59\%$, respectively. Findings suggest that highly-performing game-specific models of arousal can be obtained in a straightforward manner. How trivial would the task be across games though? The next sections are dedicated to this analysis.

### B. General Affective Models Across Games

How do we evaluate the generality of a model? The traditional way would be to test the $k$-fold cross-validation accuracies obtained within each game. Doing so, however, would only measure the generality of the affect model within the game investigated. Instead, we validate the capacity of our models to predict the arousal level across games, namely *cross-game validation* or 3-game cross-validation. In other words models are trained on two games and predict ranks in the unseen third game, which acts as the validation set.

TABLE I
Cross-game validation performance as improvement over the baseline, along with the 95% confidence interval. Significant positive improvements ($p < 5\%$) in bold.

| | | $\mu_A$ | $\Delta_A^1$ |
|---|---|---|---|
| Baseline | | $68.1\%\pm12\%$ | $52.5\%\pm1.4\%$ |
| EDA | Perceptron (NE) | $-12.9\%\pm2.0\%$ | $-2.2\%\pm9.4\%$ |
| | MLP (NE) | $-8.7\%\pm9.3\%$ | $-2\%\pm4.4\%$ |
| | RankSVM | $-20.9\%\pm11.4\%$ | $-5.7\%\pm2.9\%$ |
| GAMEPLAY | Perceptron (NE) | $0.0\%\pm12.0\%$ | **$+5.7\%\pm2.7\%$** |
| | MLP (NE) | $+0.9\%\pm11.4\%$ | $+1.5\%\pm0.6\%$ |
| | RankSVM | $-6.0\%\pm10.3\%$ | $+1.3\%\pm3.8\%$ |
| FUSION | Perceptron (NE) | $-0.4\%\pm12.7\%$ | $-1.3\%\pm9.7\%$ |
| | MLP (NE) | $-0.1\%\pm10.3\%$ | $0.0\%\pm6.0\%$ |
| | RankSVM | $-13.6\%\pm10.8\%$ | $+0.7\%\pm4.3\%$ |

Testing via cross-game validation is expected to challenge any machine learning approach. The level of variance across games is purposefully very high, making the problem of general affect modeling very difficult. In addition to the challenges of interpersonal variations of physiology, the games are very different in terms of genre, mechanics, interaction modes and gameplay experience. As a result, the arousal annotations are expected to vary with respect to all these factors.

As previously mentioned, building on the ordinal nature of subjective constructs (as emotions), our working hypothesis is that treating the arousal annotations in a relative fashion will yield more general models of arousal. To test this hypothesis, we compare between two metrics ($\mu_A$ and $\Delta_A^1$) across two modalities of input (EDA and gameplay) and their fusion. In addition, we compare them across three different preference learned models: a perceptron and a multi-layer perceptron with one layer of 10 hidden nodes (MLP)—both trained via neuroevolution—and the RankSVM method. Feature selection (SFS) was applied to all algorithms. Averaging across all combinations of games in the 3-game cross-validation process, the baseline performance is 68.1% for $\mu_A$ and 52.5% for $\Delta_A^1$. Table I shows the cross-game validation results in terms of accuracy improvement over the baseline.

*1) EDA Features:* When using only EDA features as input, the models' accuracies fail to reach the baselines, most likely due to the idiosyncratic nature of EDA. Although sub-par, the best accuracies for both annotation metrics are achieved by MLP. It is also worth noting that accuracies for $\Delta_A^1$ are closer to the corresponding baseline than accuracies for $\mu A$, pointing to a better predictive capacity of this output.

*2) Gameplay Features:* Based on Table I, using gameplay features as input seems to yield more general models of affect independently of the preference learning model used. The advantage of $\Delta_A^1$ over $\mu A$ is consistent with earlier results for the perceptron model, as it significantly improves the

baseline by 5.7%. For the MLP and RankSVM, $\Delta_A^1$ reaches accuracies slightly over the baseline. On the other hand, mean arousal reaches or slightly exceeds baseline performance only when neuroevolution is used (peceptron and MLP). RankSVM again yields the poorest model accuracies. The most frequently selected features for the $\Delta_A^1$ models are $M$, $E$ and $t_S$ as opposed to $t_S$ and $t_F$ for $\mu A$. It is likely that $t_S$ and $t_F$ are frequently selected for the $\mu A$ models since annotated arousal tended to increase (either because annotators failed to mentally register decreases in arousal or due to the games' increasingly arousing nature), making time alone (measured in $t_S$ and $t_F$) a strong predictor. On the other hand, $\Delta_A^1$ achieves the best models by considering two other gameplay features in addition to $t_S$. In other words, model generality is so far highest when considering input features in *relative* terms.

*3) Gameplay and EDA Fusion:* Finally, we fused EDA and gameplay features as input to the preference learned model. Based on Table I, most model accuracies are much higher than those with EDA features only but do not surpass those with gameplay input alone. EDA features not only fail to improve the general capacity of gameplay-based models but they also worsen it. That said, $\Delta_A^1$ manages to slightly pass the baseline for RankSVM. For the first time, RankSVM also outperforms NE methods with $\Delta_A^1$, implying that the RankSVM might overfit the training games less when considering $\Delta_A^1$ and both gameplay and EDA features. Various features were selected by SFS across the two annotation metrics and preference learning methods but the most predominant appear to be $t_S$, EDA mean, and average phasic driver for $\mu A$ and $G^-$, $E$, EDA standard deviation, and integral of the phasic driver for $\Delta_A^1$.

It should be noted that different folds have very different accuracies based on the confidence intervals of Table I. Generally, baseline accuracies for $\mu_A$ are far more varied (as high as $78.3\%$ when predicting SS). Rarely did any predictive model surpass the baseline for $\mu_A$ on any fold; the most successful was the MLP using gameplay features, which passed the baseline when predicting SS and Sonancia. On the other hand, $\Delta_A^1$ generally had smaller deviations, and more models passed the baseline for this relative metric. Using gameplay features, the perceptron passed the baseline in all folds, while the RankSVM and MLP passed the baseline in two folds. Even with a fusion of inputs, both the perceptron and the RankSVM pass the baseline in two folds; the perceptron even passed the baseline in two folds using EDA features alone. It is obvious that some games (folds) are easier to predict than others, and the low average performance of the perceptron with EDA or combined features is due to a single game (SS in both cases).

### C. Further Experimentation

To further assess how the annotation metrics impact the generality of cross-game models, experiments were performed with a longer reaction lag ($l = 1$) and deriving ranks with two previous windows (instead of one). In both cases, gameplay features and NE again yield the most general models (using an MLP). Furthermore, most $\Delta_A^1$ models (using gameplay and/or EDA features) perform 1%-5% over the baseline whereas only one $\mu A$ model does, corroborating the superiority of $\Delta_A^1$ over $\mu A$ in building cross-game general affective models.

## VIII. Discussion

Experiments in this paper tested how different ways of processing affect annotation (output) impact the general capacity of affect models across very different games and elicited experiences. This was intended to be a challenging task, to test the limits of the machine learning algorithms but more importantly the output used for training. While training game-specific models of affect was not challenging, the difficulty of the general affect modeling task was verified by our experiments. One of the more powerful machine learning approaches (SVM) tended to overfit to the two games it was trained on (its training accuracies reached more than 80% for most cases) and could not generalize well to the unseen game. In contrast, a simple perceptron which did not reach as high training accuracies managed to significantly exceed the baseline in cross-game validation when using gameplay inputs, although it marginally reached the baseline in most other cases. However, all results share a common pattern: processing annotations in a relative fashion (via the average gradient) is a more powerful approach for yielding general affect models compared to the absolute (mean) values.

Looking into future work, the obvious next steps have to explore other general input and output modalities, as well as expand the corpus of annotation and play data in these and other games. In terms of general input, it is obvious that gameplay features (as the best predictors of annotation in the current experiments) could be expanded to include more information, including splitting $G^+$ and $G^-$ into more categories.[6] It should be noted that SFS as a greedy approach of feature selection often failed to select game features which did not increase performance when added on their own but could do so if combined with other 'complementary' game features (for instance $G^+$ and $G^-$); testing other feature selection methods, such as backwards feature selection or evolutionary feature selection may improve the model's accuracy. As a final note, general features could be extracted from game-specific features (unique to each game) via e.g., transfer learning as in [18]; more ambitiously, such general game features could be extracted via computer vision applied on the video playthrough itself, via for instance deep preference learning [33]. In terms of the general output, other ways of processing the annotations of arousal could be explored, such as the amplitude of the arousal trace in each window or its integral [26]. Moreover, other time windows may capture arousal in a better way, e.g., by allowing longer time windows with overlap, or basing time windows on game events [26] such as those captured by $G^-$ or $G^+$. Finally, increasing the number of games played and annotated with even more dissimilar game genres could help filter out certain tendencies in the data (e.g., increasing arousal over time in survival genres), and allow for an even more

[6]For instance, death events and enemy collisions in *Sonancia* do not have the same affective response (the former is far more aggravating).

rigorous experimental protocol for cross-game validation using e.g., 4 games for training and 1 for testing, or more ambitious combinations such as 3 games for training and 2 for testing.

Our results largely validate our hypothesis within the games domain. Keeping in mind the ultimate goal of acquiring affect models that generalize across both tasks and domains, however, our hypothesis should also be tested within other domains. Our positive initial results, coupled with the evidenced benefits of ordinal approaches in affective computing [8]–[12], [26], leave us optimistic that similar findings may emerge.

## IX. Conclusion

This paper tested how self-reported continuous annotations of arousal can be mapped to gameplay and physiological features across games. A total of 31 game videos of three dissimilar digital games were annotated by the corresponding players in terms of arousal using an intuitive wheel-like interface. These arousal annotations were then converted into ranks of arousal values (mean) and changes (average gradient) between adjacent time windows in the same annotation trace. Preference learning algorithms attempted to find the best mapping between ranks of arousal and gameplay features shared among all three games, skin conductance data collected during play, or both. Results show that while constructing a model of arousal within a single game is straightforward and accurate, when it comes to using a model trained on two games to predict an unseen third game, the models often fail to surpass the baseline. However, when using a *relative* approach to capture annotation data (via the average gradient) rather than an absolute approach (via the mean), simple neural networks manage to significantly surpass the baseline when the model considers the gameplay data alone. These results support our hypothesis that a relative form of output in an affect model yields more general models within the same domain. Future work on testing this hypothesis on other domains than games is hoped to further generalize our findings across domains.

## References


[1] L. G. Humphreys, "The construct of general intelligence," *Intelligence*, vol. 3, no. 2, pp. 105–120, 1979.

[2] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artificial intelligence*, vol. 33, no. 1, pp. 1–64, 1987.

[3] J. Togelius and G. N. Yannakakis, "General general game AI," in *Proceedings of the Computational Intelligence and Games Conference*. IEEE, 2016.

[4] A. Zook, "Game AGI beyond characters," *Integrating Cognitive Architectures into Virtual Character Design. IGI Global*, pp. 266–293, 2016.

[5] C. Blair, "How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability," *Behavioral and Brain Sciences*, vol. 29, no. 2, pp. 109–125, 2006.

[6] D. Wechsler, "Non-intellective factors in general intelligence." *The Journal of Abnormal and Social Psychology*, vol. 38, no. 1, 1943.

[7] G. N. Yannakakis and A. Paiva, "Emotion in games," *Handbook on Affective Computing*, pp. 459–471, 2014.

[8] G. N. Yannakakis and H. P. Martínez, "Grounding truth via ordinal annotation," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2015.

[9] ——, "Ratings are overrated!" *Frontiers in ICT*, vol. 2, p. 13, 2015.

[10] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2017.

[11] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, 2014.

[12] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016.

[13] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player modeling," in *Dagstuhl Follow-Ups*, vol. 6. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[14] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Transactions on System, Man and Cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[15] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 54–67, March 2010.

[16] H. P. Martínez, M. Garbarino, and G. N. Yannakakis, "Generic physiological features as predictors of player experience," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2011.

[17] N. Shaker, M. Shaker, and M. Abou-Zleikha, "Towards generic models of player experience," in *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015.

[18] N. Shaker and M. Abou-Zleikha, "Transfer learning for cross-game prediction of player experience," in *Proceedings of the Computational Intelligence and Games Conference*, 2016.

[19] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[20] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proceedings of the ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[21] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. IEEE, 2013.

[22] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," *SEMAINE Report D6b*, 2010.

[23] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, 2016.

[24] Y. Knight, H. P. Martínez, and G. N. Yannakakis, "Space maze: Experience-driven game camera control," in *Proceedings of the Foundations of Digital Games*, 2013.

[25] P. Lopes, A. Liapis, and G. N. Yannakakis, "Sonancia: Sonification of procedurally generated game levels," in *Proceedings of the 1st computational creativity and games workshop*, 2015.

[26] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2017.

[27] C. Holmgård, G. N. Yannakakis, H. P. Martinez, and K.-I. Karstoft, "To rank or to classify? Annotating stress for reliable PTSD profiling," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2015.

[28] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[29] J. Fürnkranz and E. Hüllermeier, *Preference learning*. Springer, 2010.

[30] G. N. Yannakakis, "Preference learning for affective modeling," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2009.

[31] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the international conference on Knowledge discovery and data mining*, 2002.

[32] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," *arXiv preprint arXiv:1506.01709*, 2015.

[33] H. P. Martínez and G. N. Yannakakis, "Deep multimodal fusion: Combining discrete events and continuous signals," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014.