KI - Künstliche Intelligenz (2020) 34:45–55
https://doi.org/10.1007/s13218-020-00641-2

TECHNICAL CONTRIBUTION

I Feel I Feel You: A *Theory of Mind* Experiment in Games

David Melhart[1] · **Georgios N. Yannakakis**[1] · **Antonios Liapis**[1]

Received: 14 November 2018 / Accepted: 23 January 2020 / Published online: 4 February 2020

Abstract
In this study into the player's emotional *theory of mind* (ToM) of gameplaying agents, we investigate how an agent's behaviour and the player's own performance and emotions shape the recognition of a frustrated behaviour. We focus on the perception of frustration as it is a prevalent affective experience in human-computer interaction. We present a testbed game tailored towards this end, in which a player competes against an agent with a frustration model based on theory. We collect gameplay data, an annotated ground truth about the player's appraisal of the agent's frustration, and apply face recognition to estimate the player's emotional state. We examine the collected data through correlation analysis and predictive machine learning models, and find that the player's observable emotions are not correlated highly with the perceived frustration of the agent. This suggests that our subject's ToM is a cognitive process based on the gameplay context. Our predictive models—using ranking support vector machines—corroborate these results, yielding moderately accurate predictors of players' ToM.

Keywords Theory of mind · Affective computing · Digital games · Artificial agents · Preference learning

1 Introduction

Understanding how we recognise and feel about artificially simulated emotional behaviour is central to the design of believable characters featured in modern, narrative-heavy AAA games and the research of emotional modelling and affective computing. Arguably, it is generally complex to unravel how we feel other actors (humans or agents) feel. It is also largely unknown how we represent others' emotional and cognitive patterns according to the fundamental process known as the *theory of mind* (ToM) [24, 38, 40]: the *feeling of how others feel*.

Traditionally, the ToM refers to the mental models we form about others' higher order beliefs. However, recent studies shed light on the emotional components of ToM [38, 45] as well. Throughout this paper we use a taxonomy of cognitive and emotional representation, which relies on the belief-order attribution hierarchy [37]. According to this taxonomy we refer to our own beliefs and feelings as *zero-order representation* and our mental model of another actor's beliefs and feelings as *first-order representation*. In this regard, a *second-order representation* would be another actor's recognition of our own judgement (i.e. "it knows I know its state"). However, here we focus only on the players' recognition of emotion: specifically, their first-order representation of the agent's frustration.

We argue that modelling reliably a user's ToM can be viewed as the holy grail of not just user research and user experience design, but also adaptive and creative computation for any task that involves user-agent interactions. In games, modelling ToM could revolutionise adaptivity and personalisation—e.g. in the form of dynamic difficulty adjustment, procedural content generation, interactive narrative etc.—as our knowledge about the player's understanding of game agents would afford us more nuanced control over the experience [51].

We explore the player's emotional ToM from both statistical and predictive modelling, investigating how the gameplay context and the player's emotional state during play affect their assessment of the agent's behaviour through two different lens. We process our metrics (both input and output) in an ordinal fashion, accounting for both absolute (i.e mean values) and relative (i.e. range of fluctuation) measures

✉ David Melhart
david.melhart@um.edu.mt

Georgios N. Yannakakis
georgios.yannakakis@um.edu.mt

Antonios Liapis
antonios.liapis@um.edu.mt

[1] Institute of Digital Games, University of Malta, Msida, MSD 2080, Malta

[29]. To conduct our experiments, we introduce the MAZING testbed game, in which the player competes against an artificial agent designed to exhibit frustrated behaviour based on a top-down model inspired by the theory of *computer frustration* [5]. We focus on frustration as one of the most prevalent and context-dependent affective outcomes of human-computer interaction. The main goals of our study are (a) to investigate the relationship between the gameplay context, manifestations of player emotion, and the first-order representation of the perceived frustration of the agent based on its behaviour, while (b) to explore different ways of processing the self-reported ToM.

This paper is novel to the field of games user research and affective computing as it introduces a player-centred approach to ToM in human-agent interaction. To the best of our knowledge, this is the first time ToM is examined within human-agent scenarios, where the focus is not on the model of the agent per se but rather ~~on the players' first-order affective ToM process~~. Although most studies conceptualise ToM as a highly cognitive construct, we focus on the emotional component of the process and attempt to shed light on the ways players perceive how emotional agents feel.

2 Theoretical Background

This section provides the theoretical basis for our study and the agent model. We introduce the processes behind cognitive and affective aspects of ToM and present the theory of *computer frustration* which inspired our top-down frustration model of the gameplaying agent.

2.1 Theory of Mind and Emotions

As briefly mentioned in Sect. 1, the ToM is the concept of high-level mental models. Although traditional views focused on the representation of cognitive processes [24], the concept has been recently extended with an affective dimension [38, 45]. ToM plays a central role in social cognition and interaction [21] as it enables humans to hold and manage prevalent representations of other actors, their beliefs, emotions, and cognitive processes.

ToM has been investigated from the late '70s [7, 40] and in the context of autonomous multiagent interaction from the mid-90s [1]. However, it is only recently being considered in game design and game user research. Although the bulk of studies focus on agent-based ToM modelling [3, 13], other venues consider player-player interactions [23, 26, 33] and player-game involvement [6]. Motivated by the lack of a human-agent interaction perspective, this paper explores cognitive and emotional manifestations of a human's ToM while interacting with a game agent. While traditionally ToM is concerned with beliefs, trait judgements and strategic decisions [44], we follow Damasio's *somatic marker hypothesis* [12] and approach ToM from an emotion-centric perspective.

Based on neuroscientific evidence, we differentiate between a cognitive and an affective ToM. *Cognitive ToM* is focusing on belief and knowledge representations, while *affective ToM* processes are involved in the representation of emotions [46]. However, these processes are not mutually exclusive [45]. Cognitive ToM is generally associated with brain regions involved in autonomic responses and a choice-selection downstream of the decision making process [20, 48]. Meanwhile, affective ToM involves additional areas tied to the affective and cognitive regulation of decision making processes as described in the somatic marker hypothesis [12, 14]. Evidence also shows that affective ToM relies on cognitive empathy, which is the understanding of others' emotions, and to a lesser degree on emotional contagion, a form of emotional mimicry [45]. This suggests that while it is possible to represent other actors' mental states cognitively, affective processes impact the formulation and regulation of such mental models.

The state of the art research in virtual agents, inferring goals and recognising false beliefs, is paving the way in developing bottom-up solutions for modelling artificial ToM [41]. Such approaches, however, generally do not consider modelling affective aspects of ToM [3, 13, 41]. Adopting the typology of Ref. [41] to human players, we focus on agent-specific ToM—as opposed to general ToM, which stipulates a general predictive system—and turn our investigation towards how players formulate the cognitive and affective components of ToM with regards to game-playing agents.

2.2 Computer Frustration Theory

This explorative first study of player-agent ToM addresses perceptions of frustration. Frustration is one of the most common complex affective responses experienced during human-computer interaction [5], with distinctive cognitive and behavioural patterns [9]. The model we use for our game agents relies on the principles of the *computer frustration* theory [5] which is based on the work of Refs. [2, 25]. Computer frustration is a complex model which incorporates pre-emotional appraisal, immediate emotional response, and long lasting mood [5]. Computer frustration is positioned within the information processing theories of cognition and emotion [10, 36, 42], by emphasising its role in pre-emotional appraisal.

According to the theory, frustration is triggered by the lack of anticipated change and manifests as *non-specific arousal* in the information processing system, leading to an eventual cognitive performance dysfunction. Computer frustration differentiates between incident, session, and post-session frustration and focuses on self-efficacy, appraisal, and

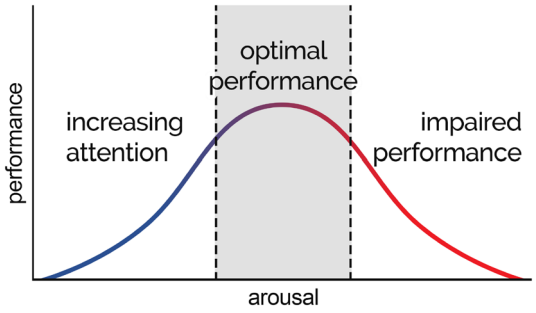

Fig. 1 Optimal level of human performance based on Ref. [25]

emotional outcomes of human-computer interaction. However, given the fast-paced nature of the game we designed for testing our hypotheses, in this paper we concentrate on the *short-term* effects and functions of frustration. Computer frustration further predicts that the severity of the interruptions and the time lost are the primary causes of incident level (moment-to-moment) frustration—whereas low self-efficacy and negative mood have a greater effect on session level and post-session outcomes.

However, not all frustrating events are detrimental to one's performance. Instead, computer frustration posits a bell-curve-like function between the level of arousal and performance [5], based on a Hebbian interpretation (Fig. 1) of the Yerkes–Dodson Law [54]. Due to the connection between arousal and performance, frustration initially has a positive effect by limiting peripheral processes (both in perception and information processing), and thus helps focus on the task at hand. This enhancing effect is especially true if the frustration originates from unmet goals or expectations (*in-game frustration*) rather than from a failure to operate an input device (*at-game frustration*) [22].

3 The MAZING Game

To collect data on a game featuring an artificial agent that might exhibit frustration, we developed a 2D top-down shooter game where a player and an artificial agent compete (Fig. 2). The player scores points by attacking the agent, while avoiding it. A game session automatically ends after 1 min.

In this study, we collect data from four playthroughs per player: in each playthrough, the opponent is different in terms of its level of frustration. The first agent has no integrated model of frustration (the value of frustration remains at 0). The other three agents are reactive to their environment and vary their frustration scores according to our model between 25–50, 50–75, and 75–100. In the following sections we detail the player's and agent's goals in this game.

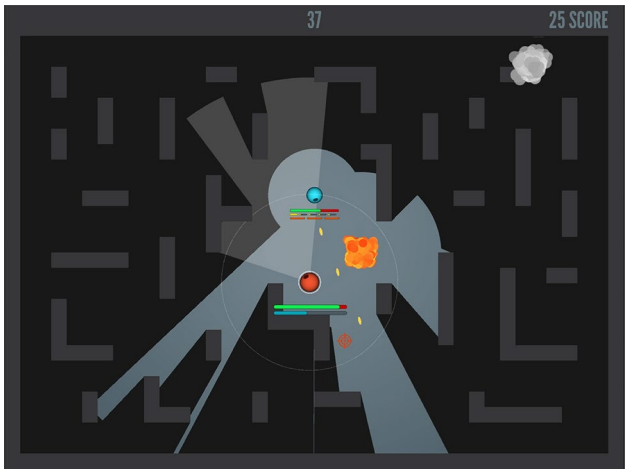

Fig. 2 Screenshot of *MAZING*, showing the player attacking the agent (teal and red orbs) and a fire in the middle. A previously laid fire is disappearing in an upper corner

3.1 Player's Mechanics

Players move in a 2D maze (viewed from a top-down perspective) using the *WASD* keys and aiming with the mouse. Their movement speed is higher than the agent's base speed, giving the player an upper hand in most scenarios. They can also use a short *dash* ability every 2 s, which grants them a speed multiplier for less than a second. ~~The player scores points by damaging the agent, via two modes of attack~~: (a) shooting up to five projectiles in quick succession by holding down the left mouse-button, and (b) throwing bombs with the right mouse-button. Bombs create fires where they land for 5 s (see Fig. 2). Passing through fire carpets deals damage to both the player and the agent, and agents are generally discouraged from moving through them. Both attacks recharge after a short period of time. If the agent dies, players gain additional points. The game obscures the maze with a partial fog-of-war, which hinders visibility but does not block it completely. Players' avatars have their field of view which illuminates the map primarily in a cone in front of them and to a lesser extent peripherally (including behind the avatar) as shown in Fig. 2. Players lose if they ~~collide with the agent~~ or ~~if they lose all their hit points~~ (players lose hit points only when passing through fire carpets). Losing decreases the player's score and re-spawns the agent and the player at their original locations.

3.2 Agent's Mechanics

The agent only performs movement and low-level decision making. The agent carries out a basic search behaviour, quasi-randomly wandering around the map. At the end of each search cycle, it picks a random point and makes its way

there, avoiding fires set by the player. If the agent senses the player, it engages in a chase. To sense the player, the agent possesses two distinct sensors mimicking visual and auditory senses. The visual sensor has an initial angle of 135° and a 10 m radius. The auditory senses affect an area around the agent (initially also 10 m), and have a low initial probability of detecting the player. If the player is standing within the sensor's reach, the agent's auditory system gradually increases the chance of detection and checks for the player every second. Intervening walls cut the auditory detection chance approximately in half. The agent takes damage from each bullet-hit and damage over time while standing in fire. The agent has many hit points, but they are not replenished over time.

3.3 Agent's Frustration Model

In order to provide the player with a quasi-believable, responsive agent, we create a model of frustration that drives the perception, movement and decision-making of the agent. Based on the theory of computer frustration (see Sect. 2.2), we regard the severity of the setback as the primary variable for increasing frustration. As the agent's primary short-term goal is to catch the player, all incidents that make it harder for the agent to do so increase its frustration. These incidents include player attacks, increasing distance from the player, and losing sight of the player. Since we conceptualize frustration as a form of arousal, we also give a light increase to the agent's frustration value whenever it spots the player. Given that we wish to model incident level frustration, we gradually decrease the agent's frustration whenever it is engaged in search behaviour.

Several stimuli from the game environment affect the agent's level of frustration. Frustration increases if the path towards a goal calculated in the previous frame is shorter than the current path (which indicates new obstacles or a player getting away) and decreases (at a lower rate) if it is longer. Frustration is increased when the agent spots the player and when the agent loses sight of the player. Third, the agent's health has an effect on the agent's frustration: frustration increases with each projectile hit. Finally, frustration slowly decreases in "resting periods", when the agent is in search behaviour. All modifiers to frustration are designed to provide players with more persistent feedback [28], and ensure that the agent is getting more frustrated throughout the session and cannot easily revert to its baseline.

The agent manifests frustration in several perceptible ways:

Sensory system Frustration causes increasingly focused attention by decreasing the angle of the agent's field of view (FoV): a frustrated agent can see further but at narrower angles, which can increase the chance to spot the player. Similarly, the area of the agent's auditory sensors is smaller as frustration rises, but the probability of hearing the player increases.

Movement On a basic level, frustration increases the agent's movement speed and rotation speed linearly. This improves the agent's performance in spotting the player initially. However, at high frustration levels it produces erratic movements; coupled with the narrower field of view, this can result in lower accuracy. Frustration also decreases the number of turns in search behaviour, simulating increasingly agitated behaviour.

Decision making Generally, the agent chooses more dangerous paths towards its goal when frustrated. The agent perceives paths through fire carpets as riskier; it is more likely to take a risky option the more health it has, or if a safe path to the player is considerably longer. Frustration affects the risk taking factor and biases the agent's behaviour towards being more reckless.

Behavioural outcomes We designed our frustration model to reflect observations in Ref. [9]. As frustration increased during play tests in that study, aggravated players took increasingly more and more risk, rushed forward, and paid less and less attention to their surroundings. In light of this research, we modify the agent's different systems to bias its behaviour towards this direction. The focused sensors, increased speed, and risk-taking behaviour is initially helpful for the agent, creating a *focused* state and modelling the increased attention of the agent. As frustration rises, the system produces distinctly *frustrated* behaviour, including hasty movements, reckless behaviour, and loss of peripheral senses. Higher frustration levels, however, lead to *rage* signified by erratic, jerky movement and an almost complete shut down of the agent's sensors. This behaviour is a natural fallout of the model but it is also in line with the *frustration-aggression hypothesis* [4].

4 Experimental Protocol

An experimental protocol was set up to collect data from each participant in a set of matchups with as diverse manifested frustration levels as possible. Each participant started with a tutorial level to get acquainted with the mechanics. After this, the participant played against an agent in four play sessions; each session was followed by a round of first-person annotation. During the four play sessions we recorded a number of gameplay metrics and players' facial features which were used to capture emotional manifestations during play. During the setup phase of the experiment, the facial recognition software was calibrated to each individual.

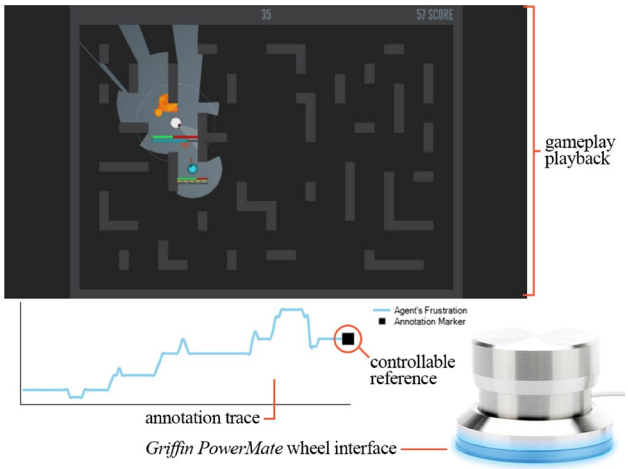

Fig. 3 The *RankTrace* software annotation tool with its physical interface

4.1 Annotation

Following the core principles of ToM, we aim to assess what our players think about the feelings of the agents they have been observing and interacting with. Players were asked to annotate the first-order representation of the agent's frustration—i.e. *how frustrated they felt the agent was*. To achieve this, the participant's last play session was recorded and played back to the player as a relived experience which the player annotated.

Labels of the agent's frustration were collected via a continuous annotation process which offers a more reliable and detailed picture of the underlying ground truth [49] and captures the temporal dynamics of the experience [30]. Specifically, the players themselves annotated their perceived frustration of the agent in every game they played. Players used the *RankTrace* tool (Fig. 3), which is an intuitive and validated [8, 29] annotation tool for unbounded and continuous annotation.

The continuous frustration trace was then converted to ordered ranks between 3-s segments of gameplay. Processing trace annotation as ordinal data provides higher reliability, generality and inter-rater agreement [8, 29, 49] and is generally better aligned with the relative nature of emotions [52].

4.2 Gameplay Features

We extracted 30 features in each gameplay session which measure the position, kinaesthetic and sensory attributes and internal states of the agent. We also consider the position and actions of the player, and the interactions between the player and the agent (e.g. distance between player and agent). Collected features refer to (a) the agent's internal values: *Frustration, Rotation Speed, Risk-Taking Factor, Movement Speed, Hearing Radius, Hearing Probability, FoV Radius, FoV Angle, Number of Turns in Search*; (b) agent behaviour: *Search Mode, Seeing Player, Chasing Player, Health, Distance Travelled, Taking Risky Path, Change in Rotation*; (c) player behaviour: *Distance Travelled, Shooting, Pressing Shoot on Cool-down, Mouse Movement, Health, Dash Pressed, Dash Mode, Pressing Dash on Cool-down, Change in Rotation, Bomb Dropping, Pressing Bomb on Cool-down*; or (d) gameplay context: *Score, Agent Distance From Player, Number of Fires*.

4.3 Facial Emotion Recognition

Neuroscienctific evidence suggests that autonomic responses alone might not be sufficient when it comes to measuring ToM [11, 20]. Emotional manifestations of ToM during gameplay are based on facial expression recognition and processing [34]. We extract facial features and derive high-level facial expressions via the Affdex SDK [32]. This system uses 34 facial landmarks to provide continuous feedback (with a rate of 10–30 FPS) and calculates the presence and intensity of the six basic emotions (anger, disgust, fear, joy, sadness, surprise) and contempt as well as estimates of the user's attention, engagement, and emotional valence from 14 facial action units. A total of 23 features are extracted from facial data captured during play and provided as intensity values of each expression on a scale between 0 (expressionless) to 100 (exaggerated display).

5 Data Preprocessing and Methods

This section discusses our methods for data preprocessing and presents two quantitative measures of our signals, metrics and annotations. Section 5.3 offers a short introduction to preference learning, focusing on ranking support vector machine (rankSVM) which is used to build predictive ToM models in Sect. 6.

5.1 Data Format and Preprocessing

Data from 80 play sessions is processed via a sliding-window approach. During this process the gameplay is segmented into consecutive equal-length windows with no overlap (w) and the mean value (μ_A) and value range ($\hat{A}$) of each feature is calculated within each window (see Fig. 4). Both μ_A and $\hat{A}$ are relevant (and disparate): the mean values are an *absolute metric* which is intuitive for comparing time windows (e.g. whether the player believes the agent is more frustrated in one window than in the next). In contrast, value range measures the amount of change in the given metrics within a time window. While value range is expressed through absolute values as well, it captures the

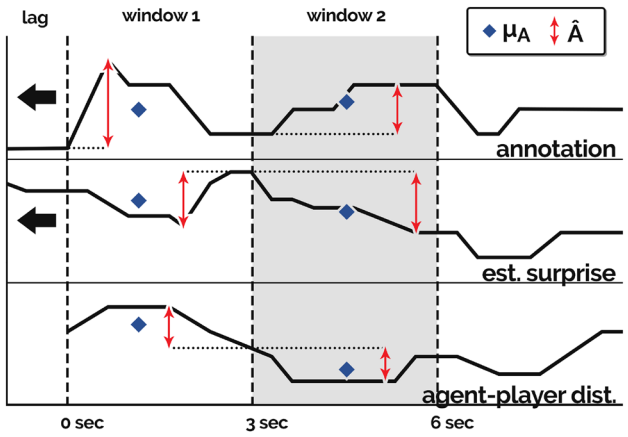

Fig. 4 Calculating the mean and value range of different signals (top to bottom: player annotation, facial data, gameplay data) through a sliding window approach. Features are shifted back 1 s in relation to the gameplay metrics before cut into equal-length windows. Mean and value range are calculated from the highlighted time window (window 2) and its previous one (window 1) to derive rankings

relative changes within a time window. Ordinal relationships of value range between time windows can be intuitive for gameplay metrics or facial expressions (e.g. whether there the game score changed more in one time window compared to the next) but, admittedly, are less intuitive for player annotation (e.g. whether the player saw a larger increase in agent frustration within one time window than in the next). Relative measures have been shown to be more powerful predictors than absolute ones for players' own affective states [8]. We believe that the degree of fluctuation within time windows can provide a clearer picture of the aggravated and erratic behaviour typical of frustrated players [9] and the fluctuation of the player's appraisal of the agent.

Based on relevant findings [30], we also consider the reaction lag of annotation traces and facial expressions (l). As in Ref. [8], in this study we parse our data with a time window of 3 s ($w = 3$), with no overlap between windows and a lag of 1 s ($l = 1$). The lag is introduced to the annotation and facial features to account for the participants' reaction time. When calculating the lag, these values are shifted back (with the first 1 s discarded) before applying the windowing method. See Fig. 4 for an illustrative example.

5.2 Method for Correlation Analysis

We use Kendall's τ for all correlation analysis reported in Sect. 6. Kendall's τ is a non-parametric, bivariate test of correlation for measuring monotonic relationships [35], which is suited for analysing the concordance of ordinal data (unlike Pearson's correlation) and is a more robust metric than Spearman's ρ but outputs lower correlation values [18]. We treat significant findings at 5% ($\alpha = 0.05$) and highly significant at 1% ($\alpha = 0.01$) level. Because multiple comparisons are being made with the same variable (the processed annotation value) the Bonferroni correction is applied. Thus, the correlation analysis measures significance at $\alpha = \frac{0.05}{53}$ and high significance at $\alpha = \frac{0.01}{53}$ for each window-processing setup (μ_A and $\hat{A}$).

5.3 Preference-Based ToM Models

Preference learning (PL) is a supervised learning technique, in which an algorithm predicts a rank order between two or more data points. The name *preference learning* originates from the most prominent applications of these algorithms in predicting user preferences [27], however, as PL simply learns to predict ordinal relationships in the data, it can be used to solve a wide array of problems where it is important to conserve the relative relation of datapoints. We use PL to investigate the ordinal change in player's emotional ToM as there is a growing body of evidence that points towards the ordinal nature of emotions [50, 52], which underlines cognitive and affective processes. Even though we focus on ordinal changes, in this paper we use the term *preference learning* to differentiate our algorithms from regression and classification algorithms. Contemporary research highlights the limitations of regression in affective computing [53]. PL is also proving more robust than classification to handle ordinal annotations of affect [8, 31, 52] as it preserves more information about the global and local order of the data than traditional class-based methods.

In this study we use a form of pairwise preference learning, which leverages binary classification by transforming the representation of the dataset from singular datapoints to pairwise differences. During this transformation each pair of input points $(x_i, x_j) \in X^2$ are observed based on their corresponding output $(y_i, y_j) \in Y^2$. Then a new dataset is constructed by assigning the pairwise difference of each pairs of input $x_i - x_j$ a label $\lambda = 1$ and $x_j - x_i$ a label $\lambda = -1$ if $y_i > y_j$ (where x_i is preferred over x_j). The resulting dataset reformulates the problem, which can be solved by any kind of binary classifier.

Because of the size of our dataset and the robustness of the technique, we use support vector machines (SVM) for this task. SVMs are supervised learning algorithms, originally designed to solve classification problems by maximizing the margin of a separating boundary between data points [47]. Since their conception, SVMs have been adapted to solve different problems including regression analysis, clustering, and—in our case—ranking [19]. In our experiments, we use the SVM implementation found in the Preference Learning Toolbox[1] [16], based on the algorithm of Ref. [27].

[1] http://plt.institutedigitalgames.com.

6 Results

Following the experimental protocol of Sect. 4, we collected data from 20 participants (described in Sect. 6.1), processing them as ranks in terms of mean values and value range of subsequent time windows. These rankings are used to analyse the impact of individual features (with rank correlations presented in Sect. 6.2) and to train predictive models which combine some or all features linearly or non-linearly (in Sect. 6.3).

6.1 Collected Data

Gameplay, facial and annotation data was collected from 20 participants (16 male, 4 female). Participants' average age was 30 and all participants held or studied towards graduate degrees. All participants were experienced players, with half of them playing daily.

Each participant played and annotated four gameplay sessions lasting 1 min each. With a sliding window of 3 s, a total of 1570 data points are collected after partially missing data was removed. These errors were caused by limitations of the face detection software. To allow participants to play freely, a web-camera was used to record their faces. As some participants shifted in their chairs during gameplay, they inadvertently moved out from the camera's vision, resulting in missing facial data.

In Sect. 6.2, 1570 individual datapoints are considered, where each datapoint represents a 3 s snapshot of a player's gameplay. For PL in Sect. 6.3, differences between all datapoints are considered. As discussed in Sect. 5.3, for each comparison two observations are made and this results in 27,968 comparisons for μ_A and 22,674 comparisons for $\hat{A}$ with a 50% baseline.

6.2 Correlation Analysis

Table 1 shows the Kendall's τ correlation values between annotated frustration of the agent and gameplay features of the agent, the player, and their interaction (i.e. General), as well as emotions estimated from facial detection. Correlations are calculated between the mean values (μ_A) of features and the annotation data, and between the value range ($\hat{A}$) of a time window for features and the annotation data. As mentioned in Sect. 5.2, we apply Bonferroni correction to all significance tests. Overall, there are only a handful of significant correlations in both μ_A (18 out of 53 with $p < 0.05$) and $\hat{A}$ (17 out of 53 with $p < 0.05$) cases. While most of the action units and more complex emotional and affective constructs measured by face recognition show very weak

Table 1 Kendall's τ correlation values between the annotation of frustration and features captured from the game and the web-cam

Type	Feature	$\tau(\mu_A)$	$\tau(\hat{A})$
Agent model	Agent Frustration Score	0.048	0.038
Agent behaviour	Search mode	**0.176**	−0.055
	Seeing player	**0.174**	**0.134**
	Chasing player	**0.169**	**0.088**
	Distance travelled	**0.125**	**0.102**
	Rotation speed	**0.101**	**0.074**
	Speed	0.048	0.035
	Change in rotation	0.008	0.016
	Taking risky path	−0.002	0.008
	Search mode length	−0.057	0.034
Agent sensory system	Health	**0.154**	**0.065**
	Hearing probability	0.054	0.033
	View radius	0.048	0.048
	Risk taking factor	0.007	0.046
	Hearing radius	−0.049	0.040
	View angle	−0.049	0.035
Player behaviour	Shooting	**0.121**	0.044
	Tries to shoot on CD[a]	**0.104**	**0.073**
	Distance travelled	**0.072**	0.026
	Mouse movement	0.029	−0.028
	Change in rotation	0.029	0.040
	Health	0.018	0.033
	Tries to bomb on CD	0.014	0.047
	Dash pressed	−0.008	−0.047
	Dashing	−0.009	−0.053
	Bomb dropped	−0.012	0.036
	Tries to dash on CD	−0.018	−0.027
General gameplay	Score	**0.240**	**0.070**
	Agent–player distance	**0.141**	**0.069**
	Number of fires	0.014	**0.071**
Basic emotions	Contempt	0.037	−0.035
	Sadness	0.017	**0.071**
	Fear	0.009	−0.058
	Surprise	0.006	0.008
	Joy	0.002	0.019
	Anger	0.001	−0.041
	Disgust	−0.018	**0.097**
Affective dimensions	Valence	**0.077**	−0.044
	Attention	−0.001	**0.090**
	Engagement	**0.070**	0.000

Table 1 (continued)

Type	Feature	$\tau(\mu_A)$	$\tau(\hat{A})$
Facial action units	ChinRaise	**0.115**	0.040
	BrowRaise	**0.064**	**0.066**
	Smirk	0.028	−0.028
	InnerBrowRaise	0.027	**0.077**
	LipSuck	0.004	−0.017
	NoseWrinkle	−0.021	**0.094**
	EyeClosure	−0.027	**0.081**
	LipPucker	−0.029	0.025
	UpperLipRaise	−0.033	**0.069**
	LipPress	−0.044	0.011
	BrowFurrow	**0.062**	−0.053
	Smile	**0.067**	0.043
	MouthOpen	**0.100**	0.057

Values in bold are significant ($p < 0.05$); highly significant values are underlined ($p < 0.01$). Bonferroni correction is applied to all significance tests

[a] *CD* cool-down. An ability is recharging and unavailable

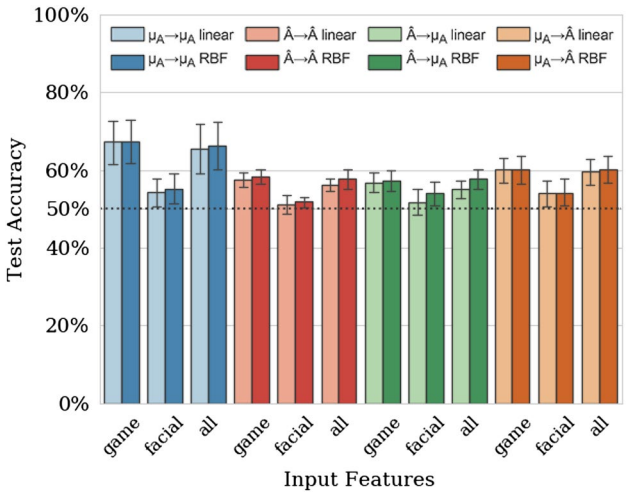

Fig. 5 Accuracies of linear SVM and best RBF SVM predictive models, on different combinations of μ_A and $\hat{A}$ input and output values (*input* → *output*). Results are averaged from tenfold cross-validation folds, and error bars denote the 95% confidence intervals

correlations (generally below 0.1), features relating to the agent's behaviour and the gameplay context show much stronger connections with the perceived frustration of the agent.

Perhaps surprisingly, the absolute highest correlation is with the player's score—which naturally relies both on the player's and the agent's performance. Even though other gameplay features inform the score (i.e. the agent's health), score is the utmost indicator of the success and failure of the player. Therefore, it provides additional high-level information about the game state compared to other, simpler features. It is also evident that captured facial features including expressions of the six basic emotions [15] and *contempt* show even weaker correlations when the data is processed as $\hat{\mu}$ and no significant connections when it is processed as μ_A. Since annotations of agent frustration have few significant correlations with affective markers but many significant correlations with contextual gameplay information, we may conclude that the first-order representation of the agents in our experiments is a predominantly cognitive process.

6.3 Predictive Models

While a traditional correlation analysis can indicate which individual features are good predictors of player ToM, it does not test how these features perform when combined in linear or non-linear fashions. We use preference learning methods (see Sect. 5.3) to construct models based on different feature sets and with the input and output processed either in an absolute or a relative fashion. The input features consist of 30 gameplay features, 23 facial emotion manifestation features, and their combination (see these 53 features in Sect. 4.2 and on Table 1). The output of our models are the ordinal relation of pairwise differences between datapoints. We infer these relation both in terms of mean values (μ_A) and value ranges ($\hat{A}$). We indicate the processing of the input and output features with a right arrow between them (i.e. in case of an input processed as mean values and output processed as value ranges the notation is $\mu_A \rightarrow \hat{A}$). To test the robustness of our models, we apply cross-participant validation: i.e. training the model on data of 18 players and testing it on data of two unseen players, repeated 10 times so that all players are validated. To measure the statistical significance of the difference between models, two-tailed t-tests are used with $p < 0.05$. When a model is tested against multiple other models, the calculation of significance is adjusted using the Bonferroni correction. Since 12 different models are compared, one model is significantly different from all others at $\alpha = \frac{0.05}{11}$.

Figure 5 shows the tenfold cross-validation accuracies of linear support vector machines (SVMs) and non-linear SVMs with radial basis function (RBF) kernels. RBF emphasises the local proximity between input vectors in a feature space, allowing for a non-linear measure of match between vectors [47]. Both linear and RBF SVMs use the C regularisation parameter to optimise the trade-off between maximising the separating margin and minimising the classification error, while the RBF SVMs also rely on the γ hyperparameter to control the weight given to datapoints during the kernel calculation. The input features and output

(annotations) are processed as mean value and value range separately, leading to four combinations of input–output. Results shown in Fig. 5 are from the best C and RBFγ values per model[2] based on an exhaustive gradient search for both the C and $RBF\gamma$ parameters from 10^{-3} to 10^{3} with powers of 10.

From Fig. 5 it is evident that the modelling of perceived frustration is a challenging predictive task. Both linear and non-linear SVMs are performing less than 10% above the baseline, with the exception of $\mu_A \rightarrow \mu_A$, which reaches 67.5% on average (80.2% at best) based on game features and 66.4% on average (81.7% at best) based on all features. Based on the results of the correlation analysis presented on Table 1, it is not surprising that features processed as $\hat{A}$ yield weaker results: the best $\hat{A} \rightarrow \hat{A}$ models only reaching 58.4% on average and 62.9% at best based on gameplay features. The best model combining both processing techniques is $\mu_A \rightarrow \hat{A}$ with 60.2% average and 69% maximum accuracy using gameplay features. While the best models are achieved using gameplay features only, accuracies for models based on facial features corroborate findings of Sect. 6.2. In the $\mu_A \rightarrow \mu_A$ and $\hat{A} \rightarrow \hat{A}$ scenarios, models based on facial recognition result in significantly worse accuracies than other feature sets. There is no significant difference, however, between models using only gameplay features and when both feature sets are combined together in a bimodal fashion.

These results are supported by Table 1 and align with the conclusion of Sect. 6.2 suggesting that the first-order representation of the agent mainly relies on the cognitive understanding of observable information without much emotional feedback. The deliberative nature of this mechanism might explain the weak predictions based on non-linear models, as players might actively interpret and reflect on the context of the interaction instead of relying on their own affective response or simple observations of the game state.

7 Discussion

This study examined the player's ToM regarding a gameplaying artificial agent which was designed to exhibit behavioural signs of frustration. The test-bed game, MAZING, was designed for the study based a contemporary theory of frustration in human-computer interaction. Within MAZING, an AI opponent was designed for the player to interact with. We collected first-person annotations of the player's first-order representation of the agent's frustration and examined the player's perception both through correlation analysis and via predictive models.

Results indicate that the most prominent correlations of the player's appraisal of agent frustration is the gameplay context, i.e. the performance and interaction of both player and agent. Our results also suggest that the process of developing and maintaining a ToM was a primarily relies on the understanding of the gameplay context, with no strong monotonic correlations to visible signs of player emotion. Predictive models of a player's ToM showed that gameplay features alone are more reliable predictors of how players appraise situations and perceive agent behaviour and frustration. On the other hand, SVM models only had moderate success in predicting players' ToM.

Our results are corroborated by Ref. [17] and recent findings of Ref. [43], which applied deep neural networks to the mapping of basic emotions to gameplay events with mixed outcomes. Just as their results, our research also indicates that the ambiguity and underlying complexity of emotions are not trivial to read and contextualise through facial emotion manifestations, leading to inaccurate predictions based on absolute measures of basic emotions. The meta-review of Ref. [39] found that multimodal modelling generally outperforms unimodal methods in audio-, video-, and text-based analysis. Our results expand these findings to new modalities which we capture through the gameplay logs (i.e. player behaviour and gameplay context) and provide additional validity by showcasing the improved performance of models using both gameplay-based, and video-based (facial features) modalities.

The primary limitation of our study is the ad-hoc nature of the agent's model of frustration: while the model is inspired by contemporary theory and manifests a varied but persistent behaviour, the testbed cannot be validated based on the statistical analysis. Preliminary comparisons between players' annotations did not show substantial differences between playthroughs with agents exhibiting low or high frustration, but future work should find a more granular method of validating the internal models of the gameplaying agent through experimentation. This could involve a focus on basic, more universally recognised emotions, a more expressive agent, and more streamlined gameplay.

Another limitation was the lack of a ground truth for the player's own emotional state, as we relied instead on detected emotion via facial expressions. While the correlation analysis showed little relationship between player emotion and perceptions of frustration, this could be instead due to the instrument used to capture emotion in the first place. We deliberately avoided an extra step asking players to annotate their own emotion, as this would cause more cognitive load and bias the ToM annotation due to ordering effects. However, future work could explore ways of collecting

[2] The best (C) and RBFγ values for each model are:$\mu_A \rightarrow \mu_A$: game: (0.1) 0.5; facial: (100) 0.1; all: (1) 0.01.$\hat{A} \rightarrow \hat{A}$: game: (0.1) 0.5; facial: (10) 1; all: (0.1) 0.1.$\hat{A} \rightarrow \mu_A$: game: (0.1) 0.1; facial: (0.1) 0.01; all: (0.1) 0.5.$\mu_A \rightarrow \hat{A}$: game: (0.5) 0.01; facial: (0.1) 0.01; all: (0.5) 0.01.

ground truth data on the emotional state of the players without increasing the difficulty of the annotation task.

Finally, while this first study focused only on gameplay metrics and facial features, future work could extend the data collection to other modalites. This study also collected a number of physiological signals (heart rate variability and electrodermal activity), but due to varying signal quality we chose to omit them from this paper. Improved ways of collecting physiological signals, gaze tracking, or other ways to process the features in a relative fashion such as average gradient per time window [8] could lead to more robust predictive models, and should be further investigated.

8 Conclusions

This paper examined a player's ToM regarding an agent's simulated frustration. The MAZING test-bed game was created explicitly towards this end, inspired by the theory of computer frustration. Results from a small-scale study with 20 players gave us a rich dataset of granular annotations of perceived agent frustration, as well as 53 features of gameplay and players' facial expressions. The analysis of the results indicated that a player's first-order representation of the agent's state is largely a cognitive process. Further, emotional responses were deemed unreliable in modelling player ToM, as relying solely on gameplay features yields models of significantly higher accuracies compared to models based on facial features.

References

1. Albrecht SV, Stone P (2018) Autonomous agents modelling other agents: a comprehensive survey and open problems. Artif Intell 258:66–95
2. Amsel A (1992) Frustration theory: an analysis of dispositional learning and memory. Cambridge University Press, Cambridge
3. Arrabales R, Ledezma A, Sanchis A (2009) Towards consciousness-like behavior in computer game characters. In: Proceedings of 2009 IEEE symposium on computational intelligence and games. IEEE, Milan, Italy, pp 217–224. https://doi.org/10.1109/CIG.2009.5286473
4. Berkowitz L (1989) Frustration-aggression hypothesis: examination and reformulation. Psychol Bull 106(1):59–73
5. Bessiere K, Newhagen JE, Robinson JP, Shneiderman B (2006) A model for computer frustration: the role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. Comput Hum Behav 22(6):941–961
6. Bormann D, Greitemeyer T (2015) Immersed in virtual worlds and minds: effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind. Soc Psychol Personal Sci 6(6):646–652
7. Bruner JS (1981) Intention in the structure of action and interaction. Adv Infancy Res 1:41–56
8. Camilleri E, Yannakakis GN, Liapis A (2017) Towards general models of player affect. In: Proceedings of 2017 seventh international conference on affective computing and intelligent interaction (ACII). IEEE, San Antonio, TX, pp 333–339. https://doi.org/10.1109/ACII.2017.8273621
9. Canossa A, Drachen A, Sørensen JRM (2011) Arrrgghh!!!: blending quantitative and qualitative methods to detect player frustration. In: Proceedings of the 6th international conference on foundations of digital games (FDG '11). Association for Computing Machinery, New York, NY, pp 61–68. https://doi.org/10.1145/2159365.2159374
10. Carver CS, Scheier MF (2012) Attention and self-regulation: a control-theory approach to human behavior. Springer, Berlin
11. Critchley HD, Corfield D, Chandler M, Mathias C, Dolan RJ (2000) Cerebral correlates of autonomic cardiovascular arousal: a functional neuroimaging investigation in humans. J Physiol 523(1):259–270
12. Damasio AR (1996) The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philos Trans R Soc Lond Ser B 351(1346):1413–1420
13. De Weerd H, Verbrugge R, Verheij B (2015) Higher-order theory of mind in the tacit communication game. Biol Inspired Cogn Archit 11:10–21
14. Dunn BD, Dalgleish T, Lawrence AD (2006) The somatic marker hypothesis: a critical evaluation. Neurosci Biobehav Rev 30(2):239–271
15. Ekman P, Freisen WV, Ancoli S (1980) Facial signs of emotional experience. J Person Soc Psychol 39(6):1125–1134
16. Farrugia VE, Martínez HP, Yannakakis GN (2015) The preference learning toolbox, p 3. arXiv:150601709
17. Fernández-Dols JM (2013) Advances in the study of facial expression: an introduction to the special section. Emot Rev 5(1):3–7
18. Fredricks GA, Nelsen RB (2007) On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. J Stat Plan Inference 137(7):2143–2150
19. Fürnkranz J, Hüllermeier E (2003) Pairwise preference learning and ranking. In: Lavrač N, Gamberger D, Blockeel H, Todorovski L (eds) Machine Learning: ECML 2003, vol 2837. Lecture Notes in Computer Science. Springer, Berlin, pp 145–156. https://doi.org/10.1007/978-3-540-39857-8_15
20. Gallagher HL, Frith CD (2003) Functional imaging of 'theory of mind'. Trends Cogn Sci 7(2):77–83
21. Garfield JL, Peterson CC, Perry T (2001) Social cognition, language acquisition and the development of the theory of mind. Mind Lang 16(5):494–541
22. Gilleade KM, Dix A (2004) Using frustration in the design of adaptive videogames. In: Proceedings of the ACM SIGCHI international conference on advances in computer entertainment technology (ACE '04). Association for Computing Machinery, New York, NY, pp 228–232. https://doi.org/10.1145/1067343.1067372
23. Goodie AS, Doshi P, Young DL (2012) Levels of theory-of-mind reasoning in competitive games. J Behav Decis Mak 25(1):95–108
24. Gopnik A, Wellman HM (1992) Why the child's theory of mind really is a theory. Mind Lang 7(1–2):145–171
25. Hebb DO (1955) Drives and the CNS (conceptual nervous system). Psychol Rev 62(4):243–254
26. Hedden T, Zhang J (2002) What do you think i think you think? Strategic reasoning in matrix games. Cognition 85(1):1–36
27. Joachims T (2002) Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). Association for Computing Machinery, New York, NY, pp 133–142. https://doi.org/10.1145/775047.775067
28. Lankoski P, Björk S (2007) Gameplay design patterns for believable non-player characters. In: Proceedings of the 2007 DiGRA international conference: Situated Play, University of Yokyo, pp 416–423. ISSN 2342-9666
29. Lopes P, Yannakakis GN, Liapis A (2017) Ranktrace: relative and unbounded affect annotation. In: Proceedings of 2017 seventh

international conference on affective computing and intelligent interaction (ACII). IEEE, San Antonio, TX, pp 158–163. https://doi.org/10.1109/ACII.2017.8273594
30. Mariooryad S, Busso C (2013) Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In: Proceedings of the 2013 humaine association conference on affective computing and intelligent interaction (ACII '13). IEEE Computer Society, USA, pp 85–90. https://doi.org/10.1109/ACII.2013.21
31. Martinez H, Yannakakis G, Hallam J (2014) Don't classify ratings of affect; rank them!. IEEE Trans Affect Comput 1:1–1
32. McDuff D, Mahmoud A, Mavadati M, Amr M, Turcot J, Kaliouby Re (2016) Affdex SDK: a cross-platform real-time multi-face expression recognition toolkit. In: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems (CHI EA '16). Association for Computing Machinery, New York, NY, pp 3723–3726. https://doi.org/10.1145/2851581.2890247
33. Meijering B, Van Rijn H, Taatgen N, Verbrugge R (2011) I do know what you think i think: second-order theory of mind in strategic games is not that difficult. Proc Annu Meet Cogn Sci Soc 33:2486–2491
34. Michel P, El Kaliouby R (2003) Real time facial expression recognition in video using support vector machines. In: Proceedings of the 5th international conference on multimodal interfaces (ICMI '03). Association for Computing Machinery, New York, NY, pp 258–264. https://doi.org/10.1145/958432.958479
35. Nelsen R (2001) Kendall tau metric. Encycl Math 3:226–227
36. Ortony A, Clore GL, Collins A (1990) The cognitive structure of emotions. Cambridge University Press, Cambridge
37. Perner J, Wimmer H (1985) "John thinks that Mary thinks that" attribution of second-order beliefs by 5-to 10-year-old children. J Exp Child Psychol 39(3):437–471
38. Poletti M, Enrici I, Adenzato M (2012) Cognitive and affective theory of mind in neurodegenerative diseases: neuropsychological, neuroanatomical and neurochemical levels. Neurosci Biobehav Rev 36(9):2147–2164
39. Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
40. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(4):515–526
41. Rabinowitz NC, Perbet F, Song HF, Zhang C, Eslami S, Botvinick M (2018) Machine theory of mind, p 21. arXiv:180207740
42. Rauterberg M (1995) About a framework for international and information processing of learning systems. In: Falkenberg ED, Hesse W, Olivé A (eds) Proceedings of the IFIP international working conference on information system concepts: towards a consolidation of views. Chapman & Hall, Ltd., London, pp 54–69
43. Roohi S, Takatalo J, Kivikangas JM, Hämäläinen P (2018) Neural network based facial expression analysis of gameevents: a cautionary tale. In: Proceedings of the 2018 annual symposium on computer–human interaction in Play (CHI Play '18). Association for Computing Machinery, New York, NY, pp 429–437. https://doi.org/10.1145/3242671.3242701
44. Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R (2015) Deconstructing and reconstructing theory of mind. Trends Cogn Sci 19(2):65–72
45. Sebastian CL, Fontaine NM, Bird G, Blakemore SJ, De Brito SA, McCrory EJ, Viding E (2011) Neural processing associated with cognitive and affective theory of mind in adolescents and adults. Soc Cogn Affect Neurosci 7(1):53–63
46. Shamay-Tsoory SG, Harari H, Aharon-Peretz J, Levkovitz Y (2010) The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. Cortex 46(5):668–677
47. Vapnik V (1995) Chapter 5 constructing learning algorithms. In: The nature of statistical learning theory. Springer, New York, NY, pp 119–157. https://doi.org/10.1007/978-1-4757-2440-0_6
48. Weilbächer RA, Gluth S (2016) The interplay of hippocampus and ventromedial prefrontal cortex in memory-based decision making. Brain Sci 7(4):15
49. Yannakakis GN, Martinez HP (2015) Grounding truth via ordinal annotation. In: Proceedings of the 2015 international conference on affective computing and intelligent interaction (ACII) (ACII '15). IEEE Computer Society, USA, pp 574–580. https://doi.org/10.1109/ACII.2015.7344627
50. Yannakakis GN, Martínez HP (2015) Ratings are overrated!. Front ICT 2:13
51. Yannakakis GN, Togelius J (2018) Artificial intelligence and games. Springer Nature, New York
52. Yannakakis GN, Cowie R, Busso C (2017) The ordinal nature of emotions. In: Proceedings of the 2017 international conference on affective computing and intelligent interaction (ACII). IEEE, San Antonio, TX, pp 248–255. https://doi.org/10.1109/ACII.2017.8273608
53. Yannakakis GN, Cowie R, Busso C (2018) The ordinal nature of emotions: an emerging approach. In: IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2018.2879512
54. Yerkes RM, Dodson JD (1908) The relation of strength of stimulus to rapidity of habit-formation. J Comp Neurol Psychol 18(5):459–482