Trace It Like You Believe It: Time-Continuous Believability Prediction

Cristiana Pacheco*, David Melhart†, Antonios Liapis†, Georgios N. Yannakakis† and Diego Perez-Liebana*
* Queen Mary University of London, London, UK
Email: {c.pacheco, diego.perez}@qmul.ac.uk
† Institute of Digital Games, University of Malta, Msida, Malta
Email: {david.melhart, antonios.liapis, georgios.yannakakis}@um.edu.mt

***Abstract*—Assessing the believability of agents, characters and simulated actors is a core challenge for human computer interaction. While numerous approaches are suggested in the literature, they are all limited to discrete and low-granularity representations of believable behavior. In this paper we view believability, for the first time, as a time-continuous phenomenon and we explore the suitability of two different affect annotation schemes for its assessment. In particular, we study the degree to which we can predict character believability in a continuous fashion through a two-player game study. The game features various opponent behaviors that are assessed for their believability by 89 participants that played the game and then annotated their recorded playthrough. Random forest models are then trained to predict believability based on ad-hoc designed in-game features. Results suggest that a discrete annotation method leads to a more robust assessment of the ground truth and subsequently better modelling performance. Our best models are able to predict a change in perceived believability with a 72.5% accuracy on average (up to 90% in the best cases) in a time-continuous manner.**

***Index Terms*—Believability, Human-Like Agents, Preference Learning, Time-Continuous Annotation, Digital Games**

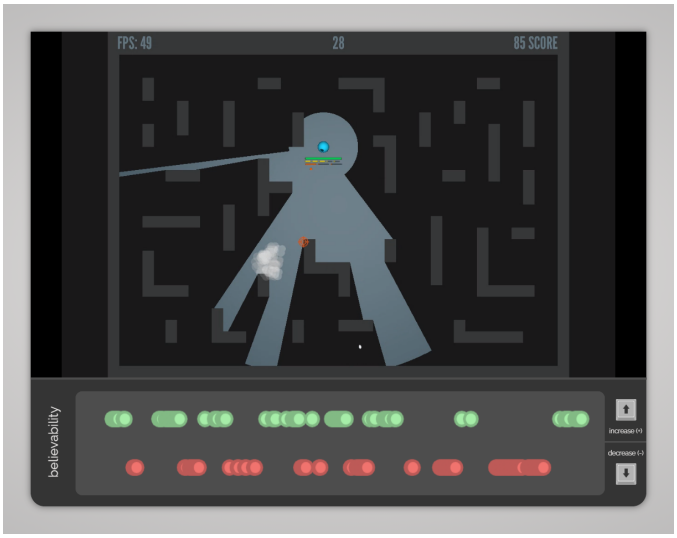

Fig. 1: Moment-to-moment believability annotation with discrete binary labels (BTrace) through PAGAN.

I. Introduction

The study of believable agents defines a core aim of affective studies and human-computer interaction at large. Not only can believable agents improve the realism of the interaction, but they can be used for human-like automatic testing [1], compete in a game like humans would do [2], [3], collaborate for solving a task [4], or guide humans through a process that may include anything from visiting a (virtual) museum all the way to a virtual therapy session. The *believability* of such agents is generally defined in terms of their behavior [5] and is predominately associated with "human-like" decisions, manifestations or expressions.

Digital games and simulated environments that feature agents are a natural testbed for studying agent believability. Within such environments one can distinguish two types of agent believability: user (or player) believability and character (or non-player) believability [5]. The former refers to the illusion that a human player is controlling the agent, rather than a computer, while the latter is applied to a fully autonomous agent which acts in a believable way to a human observer. In this study we focus on character believability in the domain of digital games.

Games put an emphasis on human-like game-playing agents, either to compete or cooperate with human players [6]. The most prominent approach for assessing believability is the adaptation of Turing Tests [7] for non-player behaviour assessment [5], [8]. However, these tests lack an established protocol and rely on an overall representation of believability. This overall representation does not always take context into account [5], [9] nor does it allow for a direct comparison between existing studies [10].

To address limitations in believability assessment, in this paper we transfer key methods and principles of affective computing to modelling believability in a *continuous* manner. We assume that features of the gameplay interaction can form efficient predictors of character believability and we test our hypothesis on an asymmetric top-down shooter game that features various non-player character behaviors [11]. We asked 89 participants to play this game and assess the believability of their opponents in a first-person annotation manner (see Figure 1). To label believability, we compare two different

This research is supported by the IEEE CIS Graduate Student Research Grants and the EP/L015846/1 for the Centre for Doctoral Training in Intelligent Games and Game Intelligence(IGGI) from the UK Engineering and Physical Sciences Research Council (EPSRC).

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

time-continuous annotation tools featured in the PAGAN annotation framework [12]: a discrete binary tool named *BTrace* inspired by *AffectRank* [13] and the continuous unbounded *RankTrace* [14]. We use preference learning via Random Forests (RFs) to investigate the robustness of models based on data collected through these two methods. Our BTrace models reach an average 72.5% accuracy (90% accuracy on the best folds) and our RankTrace models reach 68% accuracy on average (94% on the best folds). Subsequent analysis reveals that BTrace is ultimately more robust for time-continuous believability annotation. The core novelty of this study is that it approaches the assessment and modeling of character believability in a time-continuous manner. Although a continuous representation of believability has been proposed [5], to the best of our knowledge it has been untested until now. Finally, we introduce a new performance measure—$\tau@k$—which brings the statistical properties of Kendall's tau [15] to the *precision at* k ($p@k$) metric [16], [17]. $p@k$ is used in affective computing to test the comparative robustness of different algorithms. With $\tau@k$, we aim to measure the point-wise prediction performance of preference learning models trained on different believability metrics in a fair manner.

II. Background

This section reviews research related to models of simulated agents' emotion, believability of computer agents and their evaluation, and concludes with a discussion on relevant annotation protocols we can adopt from affective computing.

A. Modelling and Simulating Agent Emotion

Research into simulated emotions often relies on computer models based on *appraisal theory* [18], [19]. Appraisal theory explains emotion manifestation as a "subjective appraisal of an antecedent event" [20]. Over the years, several different frameworks have been developed to model affective processes with the goal of producing believable emotional expressions and aid the regulation of simulated behaviours. Most of these models are explicitly or implicitly based on the *Ortony–Clore–Collins* (OCC) model, which describes emotion manifestation as function of event appraisal in light of the performed action, its consequences, and the environmental context [21]. An early example of computational models taking advantage of OCC is the *EMotion and Adaptation* (EMA) framework, which models appraisal as a uniform but temporally causal process, accounting for both emotion manifestation and its impact on future behaviour [22]. In contrast, the *Fearnot AffecTIve Mind Architecture* (FAtiMA) system simulates emotions through a two-tier system comprised of a reactive (based on immediate events) and a deliberative (based on the future likelihood of success) component [18].

While these studies are focusing on modelling the cognitive process of emotion regulation, much less attention has been given to simulation of behavioural manifestations of emotion. One previous example of this research is the study of Melhart *et al.* [11] on players' emotional *theory of mind* [23] of abstract agents. In this experiment, the authors created a top-down shooter game—*MAZING*—with an adversarial agent. The agent was designed to simulate an increasingly frustrated behaviour towards the player and measured whether the players could notice this behavioural change. Instead of the OCC model, *MAZING* simulates the agent's behaviour based on the theory of *Computer Frustration* [24]. According to this theory, frustration is a reaction to a lack of anticipated change (unmet goals). As frustration often manifests as non-specific arousal, it initially increases the focus on the task at hand by limiting peripheral cognitive processes but eventually leads to performance dysfunction by overloading the information processing system [11], [24]. *MAZING* is designed to simulate this bell-curve-like function of incident level (moment-to-moment) frustration. While the agent in the game is not human-like in appearance, the game uses exaggerated behavioural and visual cues to aid emotion recognition [25]. The results of [11] highlighted the importance of the game context in the evaluation of emotional behaviour of computer agents. In this study we also use *MAZING* as our testbed game (see Section III-A).

B. Believability

While the concept of believability is often intuitively understood as something "human-like", the precise definition remains highly dependent on the context [25]. Early definitions describe the phenomenon as a *suspension of disbelief* [26] with the exact feeling of believability lying in the eye of the beholder. Bates notes the illusory nature of believability [26], where an observer unconsciously interprets the behaviour of a computer agent as human-like cognition as long as the agent does not actively try to destroy the illusion. Similarly, Loyall describes believability as an "illusion of life" which is highly dependent on the observers' expectations [27]. While these early examples offer more broad definitions attempting to capture the feeling of believability, later studies attempt to specify and categorise different aspects of the phenomenon.

Riedl and Young introduce a more objective way to assess *intentionality*—the feeling of an agent acting naturally and/or rationally—which they identify as a core component of believability [28]. Tence *et al.* define a balance between predictability and randomness, as well as exaggeration of behaviour, as facilitators of believability [25]. Tence *et al.* also note that perfectly human-like behaviour can seem less believable as nuances are lost in transmission [25].

Meanwhile, the field of game AI research discerns between player believability and character believability [5], [29], as discussed in Section I. In this paper, we are concerned with the latter, as the opponent in *MAZING* has different abilities and goals than the human player. Lankoski and Björk provide an overview of human-like believable Non-Player Characters (NPC) [30]. In their work, they define an NPC as believable if it is embodied, self-aware, has self-stated intentions, expresses emotions, has the ability to use natural language, and has *persistent traits* (i.e. maintaining a certain consistency without being overtly repetitive). While some of these criteria are

debatable—e.g. not all NPCs would necessarily need to use natural language—intentionality, consistency, and expressiveness are common aspects across all definitions [25], [28], [30].

C. Believability Evaluation

Despite several attempts to establish a normative standard for believability evaluation based both on generated criteria [31] and subjective assessment [8], [29], the majority of studies still use ad-hoc protocols. This poses a major limitation to the field. Finding normative protocols for believability assessment is important; while the common belief is that an agent's believability depends only on the AI that controls it [32], this is far from being true. In particular, Camilleri *et al.* showed that changing an agent's environment affects their perceived believability levels [33]. In their study the authors asked players to annotate the believability of a platform game play-through. Their results highlighted that the assessment was highly dependent on the configuration of different levels (i.e. number of enemies, number of gaps, placement of both and many others). Similarly, Pacheco *et al.* showed how changing multiple factors beyond the AI behaviour (such as game target audience, camera perspective, player experience, length of videos, etc.) can change the outcome of the assessment [10].

Most studies are using an evaluation method based on Turing Tests [9], opting for a high-level state-like representation of believability assigned to a whole gameplay session [5], [8]–[10]. While these methods address some of the research gaps in the field, they remain limited in a number of ways. Firstly, the vagueness of terms such as "believable" and "human-like" leaves much room for human bias during the evaluation. Both Togelius *et al.* [5] and Pacheco *et al.* [10] highlight this limitation, showing that assessment through Turing Tests is a highly reflexive process. Subsequently, participants perceive believability differently—with each participant having a different internalized definition of what "believable" is. Secondly, the diverse set of parameters used by different studies and competitions lead to results which are not directly comparable [9], [10]; this points to a need for standardized methods in believability assessment. Finally, the low granularity of Turing Tests means that the method can only be used for high-level quality evaluation of agents, but not in low-level, continuous predictions. We propose to address this challenge by increasing the granularity of the evaluation and observing how the perception of believability changes from moment to moment, thus accounting for the ambiguous dynamics of the phenomenon [5].

D. Time-Continuous Affect Annotation

Similarly to believability, affective states are fuzzy concepts which are often represented as continuous dimensions such as *pleasure, arousal*, and *dominance* [34]. Additionally, affective computing applications have been focusing on modelling changes in emotional states in a time-continuous manner [35] and several annotation tools have been developed for this purpose [12], [36]. While continuous annotation techniques are better equipped to capture the temporal dynamics of a certain experience [35], discrete (or discretised) labels can often reduce noise and increase inter-rater agreement [13]. A recent study by Melhart *et al.* [12] revealed that while discrete labels yield higher degrees of inter-rater agreement, some continuous annotation tools, such as *RankTrace* [14] are more intuitive to use.

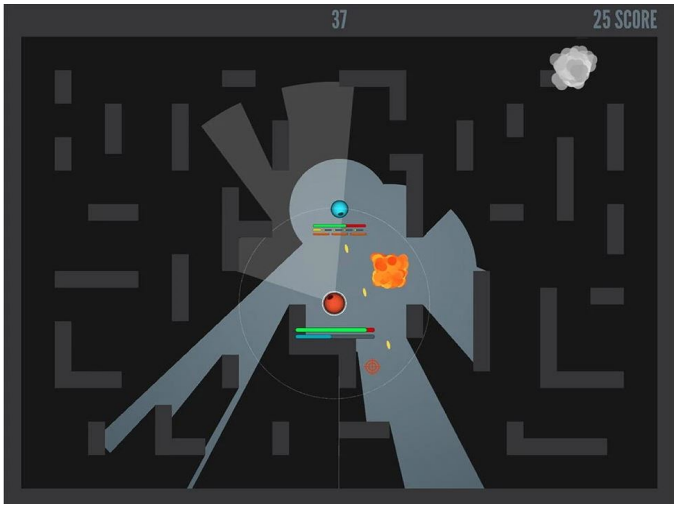

Fig. 2: Screenshot from MAZING. The blue dot is the player, the red dot is the computer agent. The light grey area shows the player field-of-vision. The image shows both player attack modes (yellow projectiles and a fire) in the middle and an extinguishing fire in the top right corner.

In this paper we adopt dominant paradigms of time-continuous affect annotation and we attempt to model character believability using both discrete traces via *BTrace* [12] and continuous traces via *RankTrace* [12], [14].

III. Believability User Study Protocol

In order to model character believability in games in a continuous manner, we conducted an online user study where participants first played a game against an artificial agent and then annotated the believability of the agent while watching a recording of their interaction. This section provides some details of the game used as elicitor for believability annotation (Section III-A), the protocol for data collection (Section III-B), and the way that data was processed to derive ordinal relationships in order to build computational models of believability (Section III-C).

A. Testbed Game: MAZING

MAZING is a top-down shooter game taking place in a maze. The goal of the player is to score points by damaging and killing a computer-controlled enemy agent. The agent reacts to the player and tries to catch them. If it collides with the player, the player dies and the game resets. A player has two options to attack: shooting fast projectiles or hurling a slower bomb which explodes into a fire that lingers on the playing field for a short period of time (see Figure 2). A player has a limited field-of-vision, obstructed by the walls

(a) RankTrace annotation

(b) BTrace annotation

Fig. 3: PAGAN data collection interfaces used for the experiments.

of the maze. MAZING is an *asymmetrical* game, and the enemy agent does not have the same abilities or statistics as the player. The enemy agent moves faster than the player, but does not have any special attacks. Instead, the enemy seeks or chases the player in order to collide with them. If the player is out of sight, the agent moves randomly through the level seeking the player; if the player is in sight the agent chases the player using the shortest path. The agent has two different sensory systems: a narrow field-of-vision, and a probability-based auditory system around the agent. While moving, the agent may pass through fires if its health is high or if the alternative path is much longer. The agents' sensorimotor skills and decision-making are affected by an abstract model of Computer Frustration [24]; more frustrated agents will take more risks, have a narrower but more precise sensory system, and move faster but more erratically.

B. Data Collection Method and Processing

We collect annotation traces about believability through the *Platform for Audiovisual General-purpose ANnotation* tool (PAGAN) [12]. PAGAN was developed as an online tool for the online collection of affect annotations. The framework was modified to house the whole data collection pipeline in one place. The *MAZING* game was integrated with PAGAN, allowing participants to play the game online and then annotate their last session. This is done in the least intrusive way possible as the online interface of PAGAN handled the assignment of play and annotation tasks automatically.

Participants were invited to play and annotate the believability of their opponent. Each player was assigned randomly between the two annotation tools: *BTrace*, a binary discrete annotation tool or *RankTrace*, an unbounded continuous annotation tool [12] (see Fig. 3). Regardless of annotation type, participants were asked to play and annotate two consecutive one-minute sessions. Before their recorded play, participants had the chance to test the controls of the game. The agents that players faced have different emotion expression capacities which were randomly assigned, based on different frustration levels (see Section III-A). After their play-session, a video replay was shown to the players and they were asked to label their opponent's gameplay in terms of believability. Before annotating, participants were informed that "believability means your opponent is playing like a human would in the situation".

For each play and annotation sequence, the collected data consists of telemetry of both player- and agent-related features and the annotation trace (continuous or binary). A sample of the telemetry data collected is included in Table II, and more details can be found in [11]. The telemetry and the annotated labels have to be resampled and aligned to each other. Game telemetry is captured at a 4Hz rate; however the 250ms time window is too small for modelling believability. Subsequently, we discretise the dataset at a 3-second interval with 1 second lag to account for annotator reaction time [37]. Consecutive 3-second time-windows are calculated based on the mean value of a given window. Finally, the features and the annotated believability values were normalized on a per-session basis.

Additional steps are taken to clean the dataset of outliers. Inspired by Makantasis *et al.* [38], the dataset is cleaned of outliers using *Dynamic Time Warping* (DTW) distance [39]. While DTW can be used to calculate a warping path between two time-series using a similarity matrix, it also provides a similarity measure in the form of cumulative DTW distance. Here, we use this metric to first measure the DTW distance to an inactive baseline; that is a hypothetical annotator who didn't label their data (all annotations are zero). We discard sessions which fall more than two standard deviations towards zero from the mean distance of the dataset to the inactive baseline. Then we calculate the cumulative DTW distance for each session compared to every other session. The goal is to identify atypical annotations and hence discard sessions which fall more than two standard deviations away from the mean cumulative distance of the dataset. Through this cleaning process, we remove believability traces with insufficient data and traces which deviated from the annotators' consensus.

C. Modelling Process

We treat both the gameplay logs and the annotations of believability in an ordinal fashion, and leverage preference learning [40] to create models of believability. We follow the extensive evidence that data treated and modelled in an ordinal fashion leads to models which are more robust compared to models constructed through a traditional classification approach [36], [41], [42]. The dataset undergoes a *pairwise transformation* of the dataset. Preference learning based on this method transforms the dataset into a new representation, which describes the relative relationship of data points and leverages binary classification to solve the machine learning task. Formally, during the pairwise transformation, for every pair of data points $(x_i, x_j) \in X$ and corresponding label $(y_i, y_j) \in Y$ we create two new data points and assign them new preference labels. In case of $y_i >_\epsilon y_j$ (x_i is preferred to x_j), we create $x' = x_i - x_j$ and $x'' = x_j - x_i$ and assign $\lambda' = 1$ and $\lambda'' = -1$ labels to them. During the pairwise comparison we use the ϵ uncertainty threshold parameter to reduce the noise in the data. Pairwise differences falling within ϵ are discarded as not significant comparisons. As an added

TABLE I: Number of participants (Part.), sessions (Sess.) and data points before and after two pre-processing treatments, split based on annotation tool used.

	RankTrace			BTrace		
Treatment	Part.	Sess.	Data	Part.	Sess.	Data
Raw	51	73	16101	38	64	14328
Preprocessed	27	40	801	24	43	860
DTW Cleaning	26	37	741	24	40	800

benefit, the baseline of the new dataset is always 50%. While in some studies the pairwise transformation is applied between every data point [43], here we apply it sequentially due to the temporal aspect of the dataset. The resulting reformulated dataset can be solved with any type of binary classifier.

After the new pairwise dataset is produced, we tackle the binary classification of ranks with Random Forests (RF). RFs are a type of ensemble machine learning model which consist of a set of randomly initialised decision trees [44]. The meta output of the RF is the *mode* of the trees' predictions. RFs are popular algorithms for the modelling of human data as they are fast to train and proven to be very robust in affective computing applications [45], [46]. In this experiment we use the RF implementation of the *Scikit-learn* Python toolkit [47]. This implementation is based on an optimised Classification And Regression Tree (CART) algorithm [48].

IV. Results

Based on the protocol described in Section III, 89 individuals participated in the user study. Participants were recruited among the authors' contacts, following a mix of purposive sampling (as participants had experience in digital games) and convenience sampling. The resulting dataset contains gameplay telemetry from a total of 89 participants, and believability annotation traces from 38 participants for BTrace and 51 participants for RankTrace. During a first preprocessing step, inactive annotators (i.e. those with fewer than 10 annotations), incomplete, or unlabelled sessions are removed. Through this process a total of 24 participants and 33 sessions are removed from RankTrace, and 14 participants and 21 sessions from BTrace. After resampling the traces and telemetry to 3 second windows (see Section III-B), we collect a total of 801 datapoints for RankTrace and 860 datapoints for BTrace. After applying the two processes for dynamic time warping to remove outliers (see Section III-B), the final dataset consists of 741 datapoints for RankTrace and 800 datapoints for BTrace. All the details of the dataset during the different cleanup phases are provided in Table I.

The following sections present the results of our analysis and modelling efforts on the collected dataset. Section IV-A discusses a correlation analysis between individual game telemetry features and annotated believability, while Section IV-B shows the results of predictive user modelling via preference learning. Throughout this section significance is reported via two-tailed Student's t-tests at $\alpha = 0.05$ and corrected with the Bonferroni method where necessary.

TABLE II: Ten absolute highest Kendall τ values between individual features and the annotated believability via BTrace and RankTrace. All correlations are significant at $\alpha = 0.05$. Labels show parameters that refer to the agent (A), the player (P), or general (G); the latter include the user's controller input.

BTrace			RankTrace		
Feature		τ	**Feature**		τ
A	Agent Chasing	0.300	A	Agent Chasing	0.194
P	Player in Sight	0.275	P	Player Score	0.153
P	Player Idle Time	-0.209	G	Time Passed	0.133
A	Agent Movement	0.209	P	Player in Sight	0.132
A	Agent-Payer Dist.	-0.198	A	Agent-Payer Dist.	-0.130
G	Input Intensity	0.192	A	Agent Movement	0.115
P	Player Score	0.187	A	Agent's x coord	0.104
G	Input Diversity	0.158	A	Agent Health	-0.098
G	Time Passed	0.152	P	Player Idle Time	-0.090
A	Agent Health	-0.149	G	Mouse x coord	0.082

A. Correlation Analysis

As a first step in our experimentation we use correlation analysis to find potential linear relationships between the perceived believability of the agent and individual gameplay telemetry features. In particular, we measure correlation through Kendall's rank correlation coefficient [49] (τ). Kendall's τ is a monotonic rank correlation measure which offers high robustness against outliers; this property makes it ideal for handling affect annotation data which is subjective and noisy by nature. We compute the τ values between all 55 in-game features and believability annotations retrieved via both RankTrace and BTrace. Due to space constraints, Table II shows only the 10 most significant correlations for each of the annotation methods.

Overall BTrace yields more and stronger linear relationships with in-game features. In particular, BTrace and RankTrace annotation traces have significant correlations with 23 and 17 out of the 55 ad-hoc designed features, respectively. Between the two annotation tools there are 14 common features that are significantly correlated to annotated believability. Most of these are related to the *agent's behaviour* (8) followed by *general gameplay* features (5), and finally features describing the *players' behaviour* (1). Specifically, the most believable aspects of the agent seem to correspond to being engaged in chasing the player, the distance travelled and whether the player is being seen or not. A possible explanation for this is that the more the agent sees and chases the player, the more interactions and focused behaviour it displays. This is also corroborated by the negative correlations between believability and the player standing idle, the agent's health, and the distance between the agent and the player. These features show that the more action happened on screen, the more likely players interpreted the agent's action as believable. This finding is also aligned with theories in believability that link it to expressivity and intentionality [25], [28].

The higher number of significant correlations and higher absolute values of τ coefficients between telemetry and BTrace annotations suggest that this tool provides annotation traces that are easier to predict through in-game contextual features

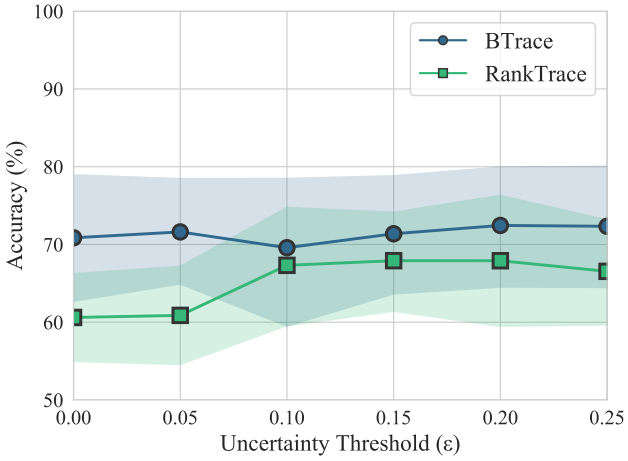

Fig. 4: Test accuracy of believability models trained on BTrace and RankTrace data over different ϵ parameters. Shaded areas show 95% confidence intervals. Baseline accuracy is 50%.

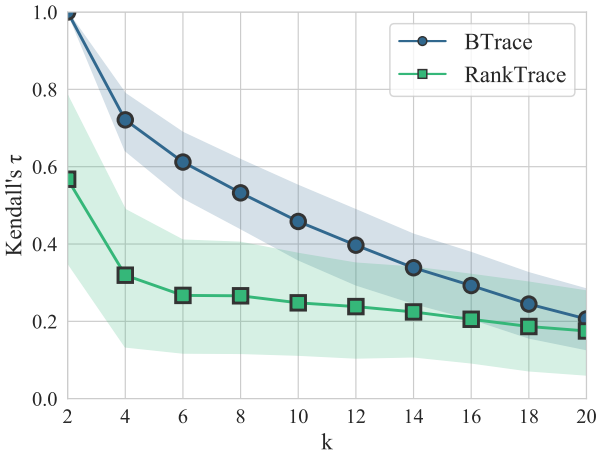

Fig. 5: Kendall's $\tau@k$ comparison between BTrace and RankTrace. The graph shows the average performance at 'k' over all test sessions. Shaded areas show 95% confidence interval.

in a linear fashion.

B. Preference Learning

This section discusses the results of believability modelling through preference learning (PL). We present and compare models based both on BTrace and RankTrace annotations. To provide a fair analysis on model performance and the robustness of the different annotation methods, we review the results using two different tasks. First, we observe model performance in terms of accuracy in a *pairwise* task. While performance on this task can tell us about the robustness of the models—since BTrace and RankTrace are using different measurements—a direct comparison between their performance can be biased. To remedy this, we introduce a *point-wise* prediction task using a new metric inspired precision at k metric [16], [17].

1) Pairwise Prediction Accuracy: Figure 4 shows the performance of our models on the pairwise prediction task. In this task we measure the accuracy of the models when predicting the more believable gameplay segment from two examples. Through these experiments we also test model sensitivity to the ϵ parameter. As annotation data is normalised on a per-session basis, an ϵ value of 0.1 means that only data points with a higher than 10% difference are considered during *pairwise transformation* (see Section III-C). While there are no significant differences between the two models across all parameters, BTrace consistently outperforms RankTrace over all ϵ values tested. Given how BTrace generally retains more data points when ϵ is increased, it seems to be the more robust approach to believability annotation. The best BTrace model (with $\epsilon = 0.2$) achieves a 72.5% accuracy on average, with a 90.1% accuracy on the best fold. In contrast, the best RankTrace model (with $\epsilon = 0.15$) reaches 67.9% accuracy on average, with 93.9% on the best fold.

2) Point-wise Prediction Performance: In this task, we reconstruct a global order of data points by taking the cumulative score of pairwise predictions in the test set. Similarly to Melhart *et al.* [42], we use the decision function of the binary preference learning task to retrieve a fine-grained cumulative rank score. In this case, for every data point (x), we make a comparison to every other data point ($x' \in X$), then simply multiply the predicted class (1 if x is preferred, -1 if x' is not preferred) with the prediction probability of the class, and sum these scores over all $x' \in X$. The assumption of this method is that the more comparisons a data point wins and the more confident the system is in the given prediction, the higher its rank score should be. We test the performance of this global ordering by looking at the Kendall's τ correlation with the raw ground truth labels using *Kendall's τ at 'k'*.

Kendall's τ at k ($\tau@k$) substitutes the precision metric in *precision at k* ($p@k$) with Kendall's τ rank correlation. Compared to $p@k$ this metric provides a greater resolution since it demonstrates the ways a model deviates from the ground truth beyond a simple classification of 'high' and 'low' data points over a median value split. The core steps of calculating $\tau@k$ are the same as $p@k$: for all possible k, the top and bottom $k/2$ predictions are observed and the Kendall's τ rank correlation with the underlying ground truth is calculated. We calculate $\tau@k$ on a per-session basis (i.e. 20 time windows of 3 seconds each). For k equal to the size of the dataset (20 in our case) we assess Kendall's τ on the whole dataset, while for $k < N$ we ignore the 'middle' of the dataset assuming that more extreme values manifest stronger and therefore more reliable and relevant reactions.

Figure 5 shows the results of this analysis of the best BTrace and RankTrace models. While at high k values the performance of the two tools overlaps without any significant difference, if k if smaller than 8 (i.e. focusing on the most extreme $\leq$40% of the session's windows), BTrace yields models that outperform models of RankTrace significantly

with respect to $\tau@k$. This analysis shows the robustness of BTrace over RankTrace, especially predicting the difference between critically high and low moments of believability.

V. Discussion

This paper presented a study on time-continuous character believability assessment and modelling. We investigated two annotation methods, discrete binary (BTrace) and continuous unbounded (RankTrace) labelling to find which one of these methods provides a more robust ground truth. To this end we used a testbed game, which featured an agent that could express different depths of behaviour, and collected gameplay telemetry data from 89 participants along with different ground truth annotations. Our correlation analysis of the collected features showed that both methods provide several significant correlations, including many in common to both. However, BTrace has more significant linear relationships to the features and in general these correlations are stronger than for RankTrace. In particular, strong connections have been discovered between believability and features describing the agent's behaviour or player inactivity. We also examined how robust the ground truth acquired through BTrace and RankTrace can be in a time-continuous player experience modelling task. We used preference learning to create models which predict the changes in the players' perception of believability dynamically. We examined the results from two perspectives, looking at a pairwise prediction task and a predicted global order of believability labels. Our results showed that BTrace labels are more robust in both cases as well.

To our knowledge, this is the first foray into time-continuous believability annotation methods and prediction. As a result, there is no accepted accuracy baseline. However, another contribution suggests that, within affective games, an accuracy of 70-80% is acceptable [50]. Thus, the conclusion of the study presented here supports discrete binary labelling for assessing character believability. The superior robustness of this method could be explained by the complexity of the term "believable". Instead of an emotional response, the subjective understanding of the phenomenon is highly cognitive. The discrete labels might help players reduce the noise and fuzziness that comes with the cognitive evaluation of such complex term—making BTrace a more intuitive tool to use in this scenario.

The presented study is still preliminary in nature. There is still much to be done before we can achieve a normative process of the time-continuous annotation and assessment of believability. The generality of the method is suggested by some of the findings (i.e. the strong correlation between believability and general gameplay features). This area could be further investigated by looking at a more diverse set of games and see whether the process generalises well. While the presented method provides rich data on how believability changes with time, follow-up studies should investigate whether the collected time-continuous data agrees with more traditional Turing test-like methods. From the current results it is unclear whether continuous believability annotation and Turing Tests are comparable or provide complementary information. The presented models take a simple approach, utilizing Random Forests and hand-crafted game telemetry. Future studies could explore different machine learning techniques including deep neural networks and pixel-to-believability prediction, similar to the use of gameplay pixels for arousal prediction [51]. In addition, we could revisit the results presented by this paper and help triangulate our initial finding with the use of physiological data.

Finally, the scope of this study could be expanded to many different domains. One direction could be to leverage the traces and trained models for agent design—not only within video games but also virtual reality or human-computer interaction. The feature analysis of Table II could directly inform a human designer to adapt the agent's behavior towards higher believability (e.g. enforcing a behavior that ensures that an agent keeps a player in its sights longer). More ambitiously, the believability predictions could act as rewards for a reinforcement learning agent that aims to maximize its believability potential. Another direction could focus on the assessment itself, as the time-continuous annotation could be applied to other areas such as assessing how 'human-like' is a conversation with chatbots or AI virtual assistants. With the right interface, such time-continuous annotation could be possible when interacting with physical robots. A binary annotation tool such as BTrace could easily be re-imagined as a two-button physical controller through which users can report believable (or non-believable) behaviors in real-time during interaction with a physical robot.

VI. Conclusion

In this study of character believability assessment, we introduce methods for time-continuous annotation for the first time and we investigate two different methods to annotate perceived believability of simulated behaviour in videogames. We collect data from 89 participants and, through both correlation analysis and machine learning techniques, we find the most robust descriptors of moment-to-moment believability. Our results show that a discrete binary annotation protocol provides a stronger linear predictor and leads to better performing predictive models as well. Our best models could predict a change in perceived believability with a 72.5% accuracy on average (up to 90% in the best of cases). The promising results of this first study in time-continuous annotation of believability can lead to a new way of assessing believability in more diverse sets of games and videos, as well as modelling believability based on moment-to-moment gameplay data and other modalities such as game footage.

References

[1] C. Holmgård, M. C. Green, A. Liapis, and J. Togelius, "Automated playtesting with procedural personas through mcts with evolved heuristics," *IEEE Trans. on Games*, vol. 11, no. 4, pp. 352–362, 2018.

[2] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman *et al.*, "Human-level performance in 3d multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019.

[3] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[4] C. Guerrero-Romero, S. M. Lucas, and D. Perez-Liebana, "Using a team of general ai algorithms to assist game design and testing," in *Proc. of the IEEE Conf. on Computational Intelligence and Games*. IEEE, 2018.

[5] J. Togelius, G. N. Yannakakis, S. Karakovskiy, and N. Shaker, "Assessing believability," in *Believable bots*. Springer, 2013, pp. 215–230.

[6] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[7] A. M. Turing, "Computing Machinery and Intelligence," in *Parsing the Turing Test*. Springer, 2009, pp. 23–65.

[8] P. Hingston, "A Turing Test for Computer Game Bots," *IEEE Trans. on Computational Intelligence and AI in Games*, vol. 1, no. 3, pp. 169–186, 2009.

[9] C. Even, A.-G. Bosser, and C. Buche, "Analysis of the protocols used to assess virtual players in multi-player computer games," in *Proc. of the Intl. Work-Conf. on Artificial Neural Networks*. Springer, 2017, pp. 657–668.

[10] C. Pacheco, L. Tokarchuk, and D. Pérez-Liébana, "Studying believability assessment in racing games," in *Proc. of the 13th international Conf. on the foundations of digital games*, 2018, pp. 1–10.

[11] D. Melhart, G. N. Yannakakis, and A. Liapis, "I feel i feel you: A theory of mind experiment in games," *KI-Künstliche Intelligenz*, vol. 34, no. 1, pp. 45–55, 2020.

[12] D. Melhart, A. Liapis, and G. N. Yannakakis, "Pagan: Video affect annotation made easy," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 130–136.

[13] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proc. of the Intl. Conf. on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 574–580.

[14] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2017, pp. 158–163.

[15] R. Nelsen, "Kendall tau metric," *Encyclopaedia of mathematics*, vol. 3, pp. 226–227, 2001.

[16] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proc. of the Intl. Conf. on Research and Development in Information Retrieval*. ACM, 2006, pp. 19–26.

[17] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proc. of the Intl. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 5205–5209.

[18] R. Aylett, M. Vala, P. Sequeira, and A. Paiva, "Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education," in *Intl. Conf. on Virtual Storytelling*. Springer, 2007, pp. 202–205.

[19] M. Ochs, N. Sabouret, and V. Corruble, "Simulation of the dynamics of nonplayer characters' emotions and social relations in games," *IEEE Trans. on Computational Intelligence and AI in Games*, vol. 1, no. 4, pp. 281–297, 2009.

[20] K. R. Scherer, "Appraisal theory," *Handbook of Cognition and Emotion*, pp. 637–663, 1999.

[21] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[22] J. Gratch and S. Marsella, "Evaluating a computational model of emotion," *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 1, pp. 23–43, 2005.

[23] S. M. Schaafsma, D. W. Pfaff, R. P. Spunt, and R. Adolphs, "Deconstructing and reconstructing theory of mind," *Trends in cognitive sciences*, vol. 19, no. 2, pp. 65–72, 2015.

[24] K. Bessiere, J. E. Newhagen, J. P. Robinson, and B. Shneiderman, "A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood," *Computers in human behavior*, vol. 22, no. 6, pp. 941–961, 2006.

[25] F. Tencé, C. Buche, P. De Loor, and O. Marc, "The challenge of believability in video games: Definitions, agents models and imitation learning," *arXiv preprint arXiv:1009.0451*, 2010.

[26] J. Bates, *The nature of characters in interactive worlds and the Oz project*. School of Computer Science, Carnegie Mellon University Pittsburgh, PA, 1992.

[27] A. B. Loyall, "Believable agents: Building interactive personalities." Carnegie-Mellon University, Tech. Rep., 1997.

[28] M. O. Riedl and R. M. Young, "An objective character believability evaluation procedure for multi-agent story generation systems," in *Intl. Workshop on Intelligent Virtual Agents*. Springer, 2005, pp. 278–291.

[29] D. Livingstone, "Turing's test and believable AI in games," *Computers in Entertainment (CIE)*, vol. 4, no. 1, pp. 6–es, 2006.

[30] P. Lankoski and S. Björk, "Gameplay design patterns for believable non-player characters." in *Proceedings of the DiGRA Conf.*, 2007, pp. 416–423.

[31] H. Warpefelt, M. Johansson, and H. Verhagen, "Analyzing the believability of game character behavior using the game agent matrix," in *Proceedings of the DiGRA Conf.*, 2013.

[32] J. D. Miles and R. Tashakkori, "Improving the believability of non-player characters in simulations," in *Proc. of the Conf. on Artificiel General Intelligence*. Atlantis Press, 2009.

[33] E. Camilleri, G. N. Yannakakis, and A. Dingli, "Platformer level design for player believability," in *Proc. of the IEEE Conf. on Computational Intelligence and Games*, 2016.

[34] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[35] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. of the automatic face and gesture recognition Conf.* IEEE, 2013, pp. 1–8.

[36] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Trans. on Affective Computing*, 2018.

[37] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 85–90.

[38] K. Makantasis, A. Liapis, and G. N. Yannakakis, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Trans. on Affective Computing*, 2021.

[39] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, USA:, 1994, pp. 359–370.

[40] J. Fürnkranz and E. Hüllermeier, "Preference learning," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 789–795.

[41] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Frontiers in ICT*, vol. 2, p. 13, 2015.

[42] D. Melhart, K. Sfikas, G. Giannakakis, and G. Y. A. Liapis, "A study on affect model validity: Nominal vs ordinal labels," in *Workshop on Artificial Intelligence in Affective Computing*. PMLR, 2020, pp. 27–34.

[43] D. Melhart, A. Azadvar, A. Canossa, A. Liapis, and G. N. Yannakakis, "Your gameplay says it all: Modelling motivation in tom clancy's the division," in *Proc. of the IEEE Conf. on Games*, 2019.

[44] R. J. Lewis, "An introduction to classification and regression tree (cart) analysis," in *Proc. of the society for Academic Emergency Medicine (SAEM) annual meeting*, 2000.

[45] D. Chatzakou, A. Vakali, and K. Kafetsios, "Detecting variation of emotions in online activities," *Expert Systems with Applications*, vol. 89, pp. 318–332, 2017.

[46] D. Hazer-Rau, L. Zhang, and H. C. Traue, "A workflow for affective computing and stress recognition from biosignals," in *Presented at the 7th Electronic Conf. on Sensors and Applications*, vol. 15, 2020, p. 30.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[48] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[49] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics*, pp. 508–510, 2007.

[50] S. McCrea, G. Geršak, and D. Novak, "Absolute and relative user perception of classification accuracy in an affective video game," *Interacting with Computers*, vol. 29, no. 2, pp. 271–286, 2017.

[51] K. Makantasis, A. Liapis, and G. N. Yannakakis, "From pixels to affect: A study on games and player experience," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2019.