The Affect Game AnnotatIoN (AGAIN) Dataset

David Melhart, Antonios Liapis and Georgios N. Yannakakis, IEEE Senior Member
Institute of Digital Games, University of Malta
Msida, Malta
david.melhart@um.edu.mt, antonios.liapis@um.edu.mt, georgios.yannakakis@um.edu.mt

Abstract—How can we model affect in a general fashion, across dissimilar tasks, and to which degree are such general representations of affect even possible? To address such questions and enable research towards general affective computing this paper introduces the Affect Game AnnotatIoN (AGAIN) dataset. AGAIN is a large-scale affective corpus that features over 1,100 in-game videos (with corresponding gameplay data) from nine different games, which are annotated for arousal from 124 participants in a first-person continuous fashion. Even though AGAIN is created for the purpose of investigating the generality of affective computing across dissimilar tasks, affect modeling can be studied within each of its 9 specific interactive games. To the best of our knowledge AGAIN is the largest—over 37 hours of annotated video and game logs—and most diverse publicly available affective dataset based on games as interactive affect elicitors.

I. INTRODUCTION

A core challenge of affective computing (AC) is the investigation of generality in the ways emotions are elicited and manifested, in the annotation protocols designed, and ultimately in the affect models created. To examine the degree to which general representations of affect are possible and meaningful, AC research requires access to corpora containing affect responses and annotations across dissimilar tasks, participants and annotators. Traditional large-scale AC datasets feature affect annotation of static images, videos, sounds and speech files within a narrow context through which affect is elicited from a particular task. However, such datasets cannot advance research in general AC, as stimuli used to elicit affect tend to be very similar. Even when the various tasks under annotation may vary, those are still limited to a very specific context—such as viewing a set of social interactions under a theme or playing sessions of the same game.

Motivated by the lack of corpora for the study of general properties of affect across tasks and participants, in this paper we introduce the Affect Game AnnotatIoN (AGAIN) dataset, which contains data from over 120 participants who played and annotated over 1,000 gameplay sessions. AGAIN features data collected from nine games spanning across three dissimilar genres, which were developed specifically for the purposes of the dataset (see Fig. 1). As shown in Table I along with game telemetry and self-annotated arousal labels, the dataset also features a video database of unique gameplay sessions with over 37 hours of in-game footage. The diverse nature of the AGAIN affect elicitors (games) provides a testbed for general affect detection in games [1], [2] and broadens the horizons for further research on general-purpose AI representations [3], [4] and artificial general intelligence.

The design and creation of AGAIN was guided by the following factors: a) accessibility, which is achieved through an online crowdsourcing framework; b) scalability: AGAIN is utilising the PAGAN online annotation framework [5] and, hence, one can easily populate the AGAIN database with more participants and annotators; c) extensibility: more affect dimensions and categories can be considered and integrated to the existing dataset through the customisable PAGAN annotation tool, and; d) generality: any additional online game or interactive session can be easily integrated to the experimental protocol of AGAIN. While at the time of writing the dataset

| Properties                  | Raw dataset | Clean dataset |
|-----------------------------|-------------|---------------|
| Number of Participants      | 124         | 122           |
| Number of Gameplay Videos   | 1116        | 995           |
| Number of Game-telemetry Logs | 1116       | 995           |
| Video database size         | 37+ hours   | 33+ hours     |
| Number of Elicitors         | 9 games (3 genres) |
| Gameplay/Video duration     | 2 min       |
| Annotation Perspective      | First-person |
| Annotation Type             | Continuous unbounded |
| Affective Labels            | Arousal     |
hosts 9 games annotated for arousal, AGAIN is designed with
all aforementioned factors in mind so that is able to host data
from more games and user modalities, considering alternative
affective labels.

The AGAIN dataset is unique in a number of ways. First,
it is the largest and most diverse publicly available affective
dataset based on games as interactive elicitors. Given the
breadth of elicitors offered, the dataset can be used for testing
specific affect models on one particular task (i.e. a particular
game) all the way to general models of affect across tasks
(game genres and games in general). Second, the dataset is
annotated with the core affective dimension of arousal, linking
dominant annotation practices in affective computing with
player modelling and game user research. Finally, it employs
a novel annotation framework [6] which captures subjective
annotations in a continuous and unbounded manner that can
be further processed as labels for regression, classification or
ordinal learning affect modelling tasks [7], [8].

The remainder of the paper is structured as follows. Section
II contextualises the dataset within the fields of affective
computing and affect modelling in games while Section III
offers a systematic review of existing audiovisual datasets
placing AGAIN within the literature of affective corpora. The
games used as the affect elicitors of AGAIN are described in
Section IV. Section V details the AGAIN dataset by describing
the protocol followed, the characteristics of the participants,
the data types collected, and the annotation framework used.
Section VI offers a detailed yet preliminary data analysis of
the dataset, and the paper concludes with Section VII

II. BACKGROUND

AGAIN is an accessible dataset offered for research in
affective computing at large and player modeling in particular.
This background section discusses the importance of arousal
within the field of affect representation (Section II-A) and
reviews studies for modelling the affect of game users (i.e.
players) in Section II-B.

A. Arousal as Affect Representation

While there are different approaches to affect representation
including categorical [9], [10], dimensional [11], and mixed
frameworks, the AGAIN dataset uses a dimensional rep-
success, showing the difficulty of creating general player models. Similarly, Bonometti et al. used high-level general features to characterise the gameplay context (such as activity count and activity diversity) to model engagement across six games published by Square Enix Ltd. [37].

III. AUDIOVISUAL AFFECTIVE DATASETS

The availability of large-scale corpora comprising affect manifestations that are elicited through appropriate stimuli is a necessity for affect modelling. Creating datasets that are annotated with reliable affect information is, therefore, instrumental to the field of affective computing at large. In this section we review representative affective corpora that rely on audiovisual elicitors and discuss the contribution of AGAIN to the current list of datasets that are enriched with affect labels. Table II presents the outcome of our survey.

We follow a systematic approach for reviewing the state of the art in affective corpora and examine the following factors that distinguish the surveyed datasets: the mode, type of the provided elicitors, the number of possible elicitor items, and overall size of the available video database (see second to fifth column of Table II), the number of participants and their recorded modalities (see columns six and seven of Table II), the annotation protocol in terms of the mode and type of the annotation (see columns eight and nine of Table II), the affective labels (see column ten of Table II), and finally the number of annotators and number of tasks each annotator had to complete (see the eleventh and twelfth column of Table II).

It is apparent from Table II that affective datasets have gradually—or over the last decade or so—drifted away from traditional induced elicitation and posed expressions, and instead turned towards soliciting spontaneous emotion manifestations. Most of these datasets have focused mainly on affect elicitation through passive (i.e. non-interactive) audiovisual stimuli (see second row of Table II). Passive audiovisual elicitors are a popular choice as they do not require any particular skill from the participants and are relatively easy to implement. In contrast, we meet datasets that make use of active elicitors involving tasks in dyads and videogames—including RELOCA [44] and player experience datasets such as PED [46] or the FUNii Database [47]. Compared to passive elicitors, these interactive tasks provide a more complex and multifaceted affective stimulus, while organically structuring the participants’ experience.

Most affective computing databases surveyed (see tenth row of Table II) capture affective dimensions such as arousal and valence, with some datasets offering labels for additional dimensions—such as dominance—and categorical labels. The surveyed datasets that have used games as affect elicitors—Mazeball [45], PED [46], and FUNii [47]—tend to be less focused with regards to the labels used and instead aim to capture more complex game-related user states such as engagement, fun or challenge. This core difference makes such player experience datasets distinctive to affective computing primarily because any lessons learned on traditional affective databases are not directly applicable to player experience datasets, and vice versa.

The affective datasets we survey appear to be rather split in terms of annotation type used. While some (e.g. DEAP [39], MANHOB-HCI [38]) opt for self annotation (first-person), many databases (e.g. RELOCA [44], SEWA [43]) use only a few expert annotators in a third-person manner. There is a clear trade-off between these approaches. First-person annotations are ideal for capturing the subjective appraisal of emotional content, while third-person annotations are better at labelling emotion manifestation through inter-rater agreement [48].

The above systematic review of the literature highlights a lack of large-scale databases implementing an active elicitation mode, using multiple elicitor types and adopting a first-person annotation scheme. While datasets using passive elicitors are generally larger, the cost associated with using active elicitation limits these datasets. As Table II shows, the size of databases featuring active elicitors generally cannot reach the standard of datasets featuring passive elicitors. The passive elicitors of these datasets, however, are also less diverse, generally limited to very similar annotation tasks. This does not advance research on general affect modelling, as researchers have to examine dissimilar datasets [2] that often comprise mismatching data collection methods and annotation tools or do not offer enough context variety (e.g. the FUNii Database [47] features two similar games from the same franchise). AGAIN addresses the aforementioned limitations by offering a large-scale corpus that is based on a set of dissimilar interactive affect elicitors that are annotated through a first-person protocol. While the dataset at the time of writing is limited to 9 games and their annotated arousal, the dataset is planned to be augmented through more affective dimensions and enriched through more games. The resulting dataset leverages the strength of active emotion elicitation while producing data in amounts comparable to databases featuring passive affect stimuli. Moreover AGAIN provides a diverse database for general affect modelling research that is not possible within any of the existing corpora.

We position AGAIN at the intersection of traditional affective computing corpora and datasets with a focus on player experience. By focusing on a core affect dimension (i.e. arousal) instead of a game-related complex emotional outcome, we aim to make AC research even more relevant for game user research, and vice versa. As games are highly interactive media, the captured data and annotations encode not merely player affect but also behaviour and game context. We focus on first-person annotations to better capture the subjective intricacies of gameplay. Finally, we choose to record continuous unbounded traces of arousal using RankTrace [6] via the PAGAN online annotation framework [5]. Such traces can be processed and machine learned in a number of ways including regression, classification and relational learning [7].
### Table II

A Survey of Affective Datasets of Audiovisual Content. A table entry is indicated with ‘N/A’ and ‘UNK’ if it is not available and unknown, respectively.

| Database          | Mode       | Elicitation | Participants | Annotation                                   | Annotators | Tasks |
|-------------------|------------|-------------|--------------|----------------------------------------------|------------|-------|
| MAHNOB-HCI [38]   | Passive    | Video       | 20 videos    | EEG, ECG, EDA, temp., resp., face and body video, gaze, audio | First-person | Discrete (9-step) | Arousal, valence, dominance, emotional keywords, predictability | 30 | 20 |
| DEAP [39]         | Passive    | Video       | 40 videos    | EEG, BVP, EDA, EMG, temp., resp., face video | First-person | Discrete (5-step) | Arousal, valence, dominance, liking, familiarity | 32 | 40 |
| LIRIS-ACCEDE [40] | Passive    | Video       | 9,800 videos | N/A | N/A | First-person | Pairwise | Arousal, valence | 1517(arousal) 2442(valence) | UNK |
| Aff-Wild [41]     | Passive    | Video       | 298 videos   | N/A | N/A | Third-person | Continuous bounded | Arousal, valence | 6-8 | 298 |
| AffectNet [42]    | Passive    | Image       | 450,000 images | N/A | N/A | Third-person | Continuous bounded, categorical | Arousal, valence, 8 emotion categories | 12 | 137,500 |
| Sonancia [18]     | Passive    | Audio       | 1280 sounds  | N/A | N/A | First-person | Pairwise | Arousal, valence, tension | UNK | 10 |
| SEWA DB [43]      | Passive, Active | Video     | 4 videos 1 task | 27 hours 17 hours | Facial landmarks, FAU, hand and head gestures | Third-person | Continuous bounded | Arousal, valence (dis)liking intensity, agreement, mimicry | 5 | 90 |
| RELOCA [44]       | Active     | Video       | 1 task       | 4 hours | ECG, EDA, face video, audio | Third-person | Continuous bounded | Arousal, valence | 6 | 23 |
| MazeBall [45]     | Active     | Game        | 1 game       | N/A | BVP(HRV), EDA, game telemetry | First-person | Pairwise | Fun, challenge, frustration, anxiety, boredom, excitement, relaxation | 36 | 1 |
| PED [46]          | Active     | Game        | 1 game       | 6 hours | Gaze, head position, game telemetry | First-person | Discrete (5-step), pairwise | Engagement, frustration, challenge | 58 | 1 |
| FUNii [47]        | Active     | Game        | 2 games      | N/A | ECG, EDA, gaze and head position, controller input | First-person | Continuous, discrete | Fun (cont.), fun, difficulty, workload, immersion, UX | 190 | 2 |
| AGAIN             | Active     | Game        | 9 games      | 37 hours | Game video, game telemetry | First-person | Continuous unbounded | Arousal | 124 | 9 |
IV. GAMES

Nine games, across three different genres, were designed and developed as affect elicitors specifically for the AGAIN dataset. We put careful consideration to create software which is aesthetically pleasing, representative of popular sub-genres of games, can be understood immediately with a basic level of game literacy, and produces a coherent and consistent dataset without the need of heavy pre-processing. To achieve this, the games featured in AGAIN are all created using the Unity 3D engine. The game genres were selected (racing, shooters, platformers) because they represent a good cross-section of the game genres and are among the most popular among gamers, but also because they have simple enough controls and clear mechanics so that players can pick them up quickly. Opposed to other genres, like role playing or strategy games, that require longer time investment and players to learn the specific mechanics, strategies and synergies, the games in the dataset relied on fast-paced genres and popular tropes to communicate the game rules as fast as possible. Specific games were designed under each genre are representative of the genre.

A. Racing

Three games represent the racing game genre, which is characterised by fast-paced driving along a given track. While racing games feature less direct interaction with opponents, players can often try to push other cars off the track or into a less favourable position. In all three games the races take place in a closed loop. The player always starts from the last position and has to fight their way up during the race. These games contain no combat mechanics but other cars and the environment can still act as obstacles or challenges. If they feel stuck, players can press the ‘R’ key to be reset to the last checkpoint. The three racing games included in the AGAIN dataset are as follows:

- TinyCars is a top-down arcade racing game (see Fig. 2a). The player’s view is isometric and the camera is at a fixed rotation. The controls are relative to the player’s car. The racetrack features no large obstacles but there is a jump-ramp and an overpass. While off-track, cars are slowed
The game was inspired by the classic arcade game *Super Cars II* (Magnetic Fields, 1991).

- **Solid** is a first-person rally game (see Fig. 2f) and plays similarly to games in the *Colin McRae Rally* series (Codemasters, 1998-2019). In this game the camera is fixed inside the car and the player’s vision is partially blocked by the steering wheel, the dashboard, and the hood of the car. To help with visibility, the UI includes a rear-mirror. The racetrack includes a large loop, in which the player has to speed up to pass through. There are no jump-ramps or other obstacles in the game. Going off-track slows the car down only a bit, making it a viable strategy to cut paths through curves.

- **ApexSpeed** is a third-person view speed racer-type game (see Fig. 2e), resembling *Wipeout* (Psygnosis, 1995)—or more recently *Redout* (34BigThings, 2016). The camera follows the player around in a 3D environment. The track is closed and the car moves forward automatically after the race starts. The racetrack features speed boosts, jump-ramps, and dangerous obstacles, which set players back to the last checkpoint. Because of the closed track, the cars cannot go off-road and the impact of collisions are reduced as well.

### C. Platformers

Finally, three games represent the platformer game genre; this genre’s gameplay focuses on traversal and often requires precision and dexterity. While platformers often feature enemies, the core goal of most platformer games is to reach the end of the level (or in some cases to go for as long as possible). Platformer games in the dataset had the most diverse control schemes, with *Endless* game requiring two keys to navigate and one to attack, *Pirates!* requiring three keys to navigate, and *Run’N’Gun* requiring five keys to navigate and one to attack. The three platforms of AGAIN are the following:

- **Endless** is a casual endless-runner game (see Fig. 2g). Rather than reaching the end of the level, the player’s goal is to stay alive for as long as possible on an endlessly looping map with randomly generated enemy placement. In *Endless*, the player can switch between two tracks or hit incoming enemies. The game also features pickups which can make the game harder (i.e. speed boost) or easier (i.e. slow down). Additionally, the player can collect coins to increase their score. On a collision with an enemy or an obstacle, the player loses score and the speed of the game is reset.

- **Pirates!** is a classical platformer (see Fig. 2h) with a gameplay which resembles *Super Mario Bros.* (Nintendo, 1985). The game focuses on traversal, especially jumping to solve light platform puzzles and collect coins to increase the player’s score. The game also features a health pickup akin to the Super Mushroom in *Super Mario Bros.* While the player has no weapons or special attacks, they can defeat enemies by jumping on their heads. On direct collision with the enemies however, the player is reset to the last checkpoint.

- **Run’N’Gun** is a shoot ’em up platformer (see Fig. 2i), imitating games like *Metal Slug* (SNK, 1996). While the goal of the game is for the player to reach the end of the level, the gameplay includes combat. The player has a health bar and a weapon, and the game features health pickups to replenish health. Enemies come in a large variety, with melee and ranged enemies, and bosses with multiple weapons and health bars. *Run’N’Gun* is the only platformer which awards the player for defeating enemies. If the player runs out of health, they are sent back to the last checkpoint.
V. AGAIN DATASET

Games in the AGAIN dataset were built for the WebGL platform and are played in a web-browser. The games were integrated into the PAGAN annotation platform [5], which allowed the large-scale crowd-sourcing of both the game playing and annotation tasks.

A. Protocol

The collection procedure took anywhere between 45 to 55 minutes. Participants were invited through Amazon’s Mechanical Turk service [3] and were compensated with 10 USD for their time. The only criterion for participation was prior purchase of videogames, in order to filter out potential subjects who might not have the game literacy required to play the games. Participants were greeted with an introduction screen (see Fig. 3), which informed them about the overall task and explained arousal as a feeling of tension, excitement, exhilaration or readiness and the opposite of boredom, calmness or relaxation. During the experiment, participants played and annotated all 9 games in the dataset. Each game (and each annotation task) take approximately 2 minutes to complete. During their play, game telemetry was collected at a rate of 4Hz and the game canvas was recorded in video format. The collection procedure was set up in an iterative-manner with participants playing for 2 minutes, then annotating their gameplay video for 2 minutes. The order of the games was randomised and this procedure was repeated until all games were played and annotated. After the experiment, participants filled in a simple exit-survey recording their biographical data and gaming habits.

B. Participants

Through the procedure presented in Section V-A, we collected data from 124 participants [4] which include 1,116 gameplay sessions (124 sessions per game) with detailed telemetry and over 37 hours of gameplay videos. Out of the 124 participants, one identified as non-binary, 43 as female, and 80 as male. Participants’ age varied between 19 and 55 years old (average of 33). Most participants were from the USA (82%); the remaining 22 participants came from Brazil (10 participants), Italy (3), Canada (2), India (2), Czech Republic (1), Germany (1), and Romania (1). Most participants identified as casual gamers (57%) or hard-core gamers (36%). Reflectively, the majority of participants (87%) were playing daily or weekly. All participants had either a PC or a gaming console or both, with the most popular platform being PC. Participants played very diverse games in their free time across different genres: from casual games through platformers, sports simulators, shooters, to role-playing games. The anonymised demographic data is included in the dataset.

C. Game Footage Videos

For realising first person annotation the gameplay footage of players had to be recorded and annotated by the players themselves. As a result the raw AGAIN dataset features 1,116 videos of around 2 mins each (i.e. over 37 hours of game footage). The video database contains more than $3 \times 10^6$ frames of video, which are recorded at 24 FPS and have a resolution of $960 \times 600$ pixels. The characteristics of the AGAIN corpus enable the use of data-hungry deep learning methods for directly mapping affect to frame pixels [3].

D. Game Context Features

In addition to the raw video game footage, AGAIN features a number of hand-crafted attributes for each game. Inspired by advances in machine learning with privileged information [52], we view telemetry data as privileged information and we include such ad-hoc features in the dataset. Fusing gameplay features with other user modalities has also been a dominant practice in game-based affective computing [53], [54]. The game context features described in this section are considered in the preliminary data analysis of the dataset in Section V.

Table III

| Genre    | Sub-Genre          | Title          | # Features |
|----------|--------------------|----------------|------------|
| Racing   | Arcade-Racing      | TinyCars       | 33         |
|          | Rally              | Solid          | 34         |
|          | Speed-Racer        | ApexSpeed      | 34         |
| Shooter  | First-Person Shooter | Heist!        | 37         |
|          | Top-down Shooter   | TopDown        | 38         |
|          | Arcade-Shooter     | Shootout       | 23         |
| Platform | Endless Runner     | Endless        | 33         |
|          | Mario-Clone        | Pirates!       | 39         |
|          | Shoot’Em’Up        | Run’N’Gun      | 47         |

[3] While 169 participants completed the data collection process, 45 participants were omitted as their experiments were incomplete (i.e. no video or annotation data) due to software or hardware error.
Table IV
THE GENERAL GAMEPLAY FEATURES OF AGAIN

| feature               | description                                                                                                                                 |
|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| time_passed           | time counted from the start of the recording                                                                                                 |
| score                 | player score                                                                                                                                 |
| input_intensity       | number of keypresses                                                                                                                         |
| input_diversity       | number of unique keypresses                                                                                                                   |
| idle_time             | percentage of time spent without input                                                                                                        |
| activity              | inverse of idle_time                                                                                                                          |
| movement              | distance travelled + reticle moved (in shooters)                                                                                              |
| bot_count             | number of bots visible                                                                                                                        |
| bot_movement          | bot distance travelled                                                                                                                        |
| bot_diversity         | number of unique bots visible                                                                                                                  |
| object_intensity      | number of objects of interest                                                                                                                  |
| object_diversity      | number of unique objects                                                                                                                      |
| event_intensity       | number of events                                                                                                                             |
| event_diversity       | number of unique events                                                                                                                       |

All AGAIN games implement the same data-logging strategy and use a similar method for recording telemetry. Games within the same genre share the same feature labels. Not all features, however, have a qualitative meaning for all games within a genre—in Heist!, for instance, players move but they are immobile in Shootout. To ease the data collection and aggregation process, when features are absent from a game they are given values with zero-variance (zeroes or ones, depending on the feature). For example, a looping racetrack is only present in the Solid game (see Figure 2b), therefore the visible_loop_count feature is always zero in the other racing games.

Table III shows the number of features we have extracted per game with the zero-variance features removed. The recorded game telemetry encodes control events initiated by the player (e.g. player_steering), player status (e.g. player_health), gameplay events outside of the player’s control (e.g. bot_aim_at_player), bot status (e.g. bot_offroad), and the proximal and general game context (e.g. bot_player_distance and pickups_visible). Gameplay is recorded at approximately 4Hz (every 250ms). Due to limitations of the Unity engine and the WebGL format, the logging rate is not consistent. To mitigate this issue, the logging script aggregates multiple ticks of the engine’s update loop and provides an average value. Due to this processing technique almost all events are represented by continuous values. For example, pickups_visible can take float values under 1 when a pickup just became visible at the end of the given 250ms window. The only features which are represented by integer values are player_death and objects_destroyed because of their sparsity.

In addition to the features enumerated in Table III, the dataset includes 14 general gameplay features. These general features are ad-hoc designed and derived from the game-specific events and are based on contemporary studies of general player modelling [2], [37]. Events which require expert evaluation of the game such as the goal-oriented and goal-opposed events of Camilleri et al. [2] are omitted from these general features of AGAIN, but may be considered as additional features. Table IV lists these features alongside their explanation.

E. Annotation

The annotation task was administered through the PAGAN platform [9], using the RankTrace annotation method [6]. PAGAN is an online annotation platform developed to be an easy-to-use software for crowdsourcing annotation tasks with a focus on one-dimensional time-continuous annotation using three different methods. RankTrace [6], an ordinal annotation framework, GTrace [55], a bounded annotation scale which gathers continuous data that can be converted to a Likert-like format, and BTrace, which is a binary annotation tool for both time-continuous and discrete annotation, inspired by AffectRank [56]. We have chosen RankTrace as our annotation framework for this dataset.

RankTrace allowed us to collect data in an unbounded fashion (see Fig. 4). This type of data is best interpreted as subjective, ordinal labels as it preserves the relative relationships between datapoints [7]. The unbounded trace means that users can always adjust their annotations higher or lower than previous values, which alleviates much of the guesswork compared to when users annotate on an absolute and objective scale [54]. The ordinal nature of the annotation follows the cognitive process of human evaluation, as it provides a trace which factors in habituation [57], anchoring bias [58], [59] and recency-effects [60].

F. Data Cleaning

To ease any subsequent analysis and future studies based on the dataset, in this section we propose a preprocessing pipeline which removes 10.8% of the dataset as outliers. AGAIN contains both the raw and the cleaned data that result from the process outlined here.
Since PAGAN only records annotations when there is a change in the signal and the Unity engine loop is affected by hardware performance, as a first step we resample the whole dataset at 4Hz to get a consistent signal. We remove duplicate values from the dataset, as well as sessions which are either too short (less than 1 minute) or too long (more than 3 minutes) due to software or technical errors during crowdsourcing. We also prune sessions which have less than 10 annotation points, assuming that the participant was unresponsive. This initial cleanup phase removes 24 sessions (2.1% of the data).

To clean the dataset further, we apply Dynamic Time Warping (DTW) to get an approximate similarity measure between traces. DTW is used in time-series analysis to measure the similarity between temporal sequences that might be out of sync or vary in speed \[61\]. DTW works by calculating a warping path between two signals using a similarity matrix and provides a useful metric which qualifies time-series in the form of a cumulative DTW distance describing the similarity to a baseline trace or other signals \[61\]. We apply both of these strategies when cleaning the dataset. As a first step, we calculate the cumulative DTW distance to an artificial flat baseline (arousal annotations at 0 in all time windows). The resulting score provides us with a similarity measure to an artificial session where the participant performed no annotation; this allows us to remove unresponsive outliers. We remove all sessions which fall more than two standard deviations closer to zero from the average cumulative distance (the left tail of the distribution). This step removes 28 additional sessions from the dataset (2.5%). Finally, we apply the cumulative DTW distance metric between each datapoint and sum up the resulting distances. This metric shows us the relative similarity of a session to every other session. We remove all sessions which fall more than two standard deviations away from the average summed cumulative distance (see Fig. 5). This step removes an additional 69 sessions (6.2%). This last step removes annotations which are too dissimilar from the general trends of participants’ annotations; we presume that either the annotation was improper or that this session’s elicitor was somehow not in line with how other players played the same game. At the end of the cleaning process, 121 sessions—including all data from 2 participants—are removed (10.8%). The cleaned dataset consists of 122 participants and 995 sessions; more details on the cleaned dataset are provided in Section \[VI\].

### Table V

| Game         | Sessions | Data ↑ (\%) | ↑ | ↓ | — |
|--------------|----------|-------------|---|---|---|
| TinyCars     | 109      | 52.75       | 543 | 461 | 3386 |
| Solid        | 109      | 53.42       | 613 | 492 | 3346 |
| ApexSpeed    | 114      | 56.10       | 607 | 462 | 3581 |
| Racing       | 332      | 162.27      | 1763 | 1415 | 10313 |
| Heist!       | 110      | 53.91       | 580 | 424 | 3479 |
| TopDown      | 115      | 56.90       | 650 | 463 | 3614 |
| Shootout     | 106      | 51.77       | 471 | 341 | 3496 |
| Shooter      | 331      | 162.57      | 1701 | 1228 | 10589 |
| Endless      | 112      | 55.11       | 559 | 438 | 3595 |
| Pirates!     | 110      | 52.26       | 625 | 534 | 3816 |
| Run’N’Gun    | 110      | 54.97       | 618 | 431 | 3521 |
| Platformer   | 332      | 162.34      | 1802 | 1403 | 10502 |
| Total        | 995      | 487.18      | 5266 | 4046 | 31204 |

Figure 5. Distribution of summed cumulative DTW distance values of each session compared to every other session. The solid line shows the average score, while the dotted lines show the first and second standard deviation. Values in the grey field (right tail) are removed during data cleaning.

### VI. AGAIN Analysis

Following the cleanup process presented in Section \[VI\], this section performs a preliminary analysis of the clean version of the AGAIN dataset focusing on patterns in the arousal annotations (see Section \[VI-A\]) and the AGAIN game context features (see Section \[VI-B\]). The section concludes with an initial set of affect modelling experiments in the AGAIN dataset (Section \[VI-C\]) that can serve as baseline for future studies with this dataset. While some games receive more aggressive data cleaning than others (TinyCars, Solid, and Shootout), overall there is an even distribution of data and sessions across genres as shown in Table \[V\].

#### A. Trends in Annotations

Figure \[6\] shows the average annotation trace as calculated by averaging values in time windows of 250 ms of all sessions’ traces. It is evident that arousal annotation tends to have an upwards tendency. This is not surprising, as most games considered are action-oriented with an ever-increasing challenge; for instance, Endless keeps increasing the speed.
of the game which evidently makes it both harder and more
arousing as time passes. Racing games (top row of Figure
9), on the other hand, tend to have arousal converging to
a maximum mean value after the first 30 seconds. This is
likely because the player is initially rushing to overtake
the opponents’ cars (players always start last); after this initial
excitement the race becomes repetitive, with players trying to
either maintain the lead or slowly catch up to the leader.

B. Trends in Game Context Features

Observing the twelve general gameplay features shared
across all nine games, one can detect some notable differences
between games. In terms of the player’s input (control), games
with more complex interaction schemes appear to have higher
input diversity and input intensity (see Table IV for details
on these features). Even accounting for the games’ different
control schemes (i.e. the number of controls the player has
available), ApexSpeed, Shootout, and Endless have the lowest
intensity (number of keypresses) and diversity (number of
unique keypresses) while Pirates! and TinyCars have the high-
est diversity. This discrepancy could point to an easier control
scheme for the former games, but it could also point to a
more frantic and engaging interaction in the latter games. The
idle time and activity features corroborate this observation, as
racing games have less idle time without keypresses (since
in two of the games the player needs to constantly press a
button to move forward). In contrast, games where participants
mainly reacted to stimuli (e.g. in Shootout players react to
opponents popping up and in Endless players jump only when
a gap or obstacle is near) featured much higher idle times. In

terms of other features, the number of bots (opponents) visible
on the screen varied wildly between games, with Tiny Cars
and Shootout having the highest number of visible enemies on
average. Perhaps due to the many enemies present, Shootout
had the highest number of events (event intensity in Table IV),
while Solid had the fewest events per time window.

In terms of comparing the general gameplay features across
games, this requires some normalisation in order to account
for both discrepancies in value ranges between games (e.g. in
terms of score) but also between players in the same game.
Following the paradigm of treating both input and output
as relative [2], the gameplay features of each time window
are normalised to the [0, 1] value range within each session.
As a result such normalised features consider the dynamics
of a single player in a given session (e.g. in which time
window the player achieved the top score of their session),
disregarding for example whether other players reached higher
scores in the same game. Since arousal is similarly a deeply
subjective notion, the player is expected to annotate arousal in
the context of their current session (e.g. whether their arousal
might increase if they start performing better than they were
performing previously in the same session). After all 4.9 · 10^5
game context data was normalised in this fashion, we applied
t-distributed Stochastic Neighbor Embedding (t-SNE) [62] to
map this data on a two-dimensional space. Figure 7 shows the
resulting data distributions. The visualised distributions offer
some important insights on the differences between games.
In particular, every game’s general features tend to exhibit
different patterns compared to the other 8 games. Moreover,
the compressed (game context) feature distributions across the
three shooter games (see middle row of the figure) appear
quite distinct from one another. In some cases, however, there
appears to be an overlap, either between games of the same
genre (e.g. all racing games) and games of different genres
(e.g. see Pirates! and Solid). Even though this type of data
visualisation cannot shed light on all possible differences
between games, it indicates that the games impact the patterns
of the data solicited from players (i.e. the context influences
the user behaviour). The t-SNE analysis also indicates that the
problem of mapping between game content and arousal seems
to be easier for some games (and game genres) than others.
The next section hosts an initial study that investigates the
potential of machine learning for deriving such a mapping.

C. Preliminary Arousal Models

In this section we provide an initial analysis of the AGAIN
dataset which aims to serve as a baseline study for future
affect modelling attempts with this dataset. In particular, we
process the clean AGAIN dataset to predict arousal. To that
end, we split the annotation traces in 3-second time windows,
and compute the mean arousal value from all data points in
that time window. Following common practice in affective
computing [2], [6], [5], we introduce a time offset of 1-
second to the annotation traces. As discussed in Section VI-B
all features (including the arousal values) are normalised on a
per-session basis.
We treat arousal modeling in AGAIN as a preference learning task \cite{7, 8, 54} and focus on predicting arousal change from a 3-second time window to the next. To reduce experimental noise from trivial changes within the arousal trace we omit all consecutive time windows between which the arousal change is less than 10% of the total amplitude of the session’s arousal value. While this 10% threshold is based on prior experiments in similar problems \cite{18, 65}, a more extensive analysis could explore the impact of the threshold value on prediction accuracy and the volume of data lost. Pairs of consecutive time windows where the mean arousal in the second time window is higher than the first (over the threshold) is labelled with a value of 1 (†); see Table \ref{V} and -1 if otherwise (⊥); see Table \ref{V}. As noted, pairs where the absolute difference in terms of arousal is below the 10% threshold are labelled as “stable” (―); see Table \ref{V} and omitted from the dataset. Table \ref{V} shows the distribution of ascending, descending and stable pairs of time windows per game.

By applying this pairwise transformation to consecutive time-windows the preference learning paradigm is reformulated as binary classification. To construct accessible and simple models of arousal, in this initial study we employ a Random Forest Classifier. A Random Forest (RF) is an ensemble learning method, which operates by constructing a number of randomly initialised decision trees and uses the mode of their independent predictions as its output. Decision trees are simple learning algorithms, which operate through an acyclical network of nodes that split the decision process along smaller feature sets and model the prediction as a tree of decisions \cite{66}. In this paper we are using the RF implementation in the Scikit-learn Python library \cite{67}. We initialise RFs with their default parameters. For controlling overfitting we set the number of estimators in the RF to 100 and the maximum depth of each tree to 10. This experimental setup is meant to provide a simple baseline prediction performance for the dataset, and thus, we are not tuning the hyperparameters of the algorithm in this paper any further.

To examine the validity of the general features discussed in Section \ref{V-D} models are constructed for each game based on three different feature sets: 1) game specific features excluding additional general features 2) general features that include only the features shown on Table \ref{V} and 3) all features combined. Due to the pairwise transformation discussed above, the baseline accuracy of all experiments is 50%. Because RFs are stochastic algorithms, we run each experiment 5 times and we report the 10-fold cross validation accuracy. Note that each fold contains the data of 10 to 12 participants and no two folds contain data from the same participant. The reported statistical significance is measured with two-tailed Student’s t-tests with $\alpha = 0.05$, adjusted with the Bonferroni correction where applicable.

Figure \ref{8} shows the performance of the RF models. Prediction accuracy varies between $58.06\%$ and $82.50\%$ across games. The results reveal that arousal appears to be easier to predict in some games (e.g. ApexSpeed, TopDown, and Endless) than others (e.g. TinyCars, Shootout, and Run’N’Gun). In the racing and platformer genres, games with fewer input options and an automatic progression system (ApexSpeed and Endless respectively) are tied to higher model performance. An explanation could be that games that have more internal structure (due to the sparsity of actions the player can take and automatic progression through the game with minimal input) present a simpler problem. An exception to this observation is Shootout, in which the controls are limited (only looking around and shooting) and enemies appearing in an ever-
increasing speed, but despite these similarities with ApexSpeed and Endless, Shootout models are struggling to reach 60% accuracy (the lowest performance across all games).

Looking at individual games across different feature sets, we observe that the general features manage to perform comparably to the specific features independently of the game tested. Game-specific features yield significantly higher performances than general features only in 4 games (TinyCars, Solid, Endless, and Pirates!). Moreover, the combination of both specific and general features yields significantly more accurate arousal models than either the game-specific or general features (or both) in 5 games: Solid, Heist!, TopDown, Endless, and Pirates!. These results demonstrate the robustness of the general features presented in Section V-D and show that there is little to no trade-off in representing the presented games in a more abstract and general manner.

The arousal model performances presented in this section highlight a number of challenges for future research. Firstly, the differences in performances between games show that the complexity of the affect modelling task is dependent on the characteristics of the elicitor and the game context. Finding new processing methods, data treatment, algorithms, and model architectures which perform equally well across different games is an open problem. Secondly, the robustness demonstrated by the general features proposed in this paper point towards the possibility of general affect modelling across games. While research has already been investigating general affect modelling in videogames [2], early results showed only moderate success. The dataset and baselines presented in this paper provide a large open source database of games with robust enough general features to continue the exploration of general affect modelling.

VII. DISCUSSION & CONCLUSION

This paper introduced a new database for affective modelling, the AGAIN dataset. AGAIN is the largest and most diverse publicly available dataset coupling gameplay context, video footage of games, and annotated affect to date. It includes a variety of interactive elicitors, in the form of nine games from three popular yet dissimilar game genres. In particular, the dataset consists of 37 hours of video footage accompanied by telemetry and self-annotated arousal labels from 1,116 gameplay sessions played by 124 participants. The motivation behind the construction of this dataset is to facilitate and further advance research on general affect modelling through a clean, large-scale, diverse (elicitor-wise) and accessible database.

While each game elicits similar playstyles across different participants, the database features unique videos with self-annotated arousal traces. AGAIN puts an emphasis on first-person annotation as—compared to a third-person annotation scheme—is expected to yield ground truths of affect that are closer to the affect experienced [7], [65], [68]. The existing in-game footage of AGAIN, however, can be used directly for third-person annotation in future studies. Regardless of the annotation scheme used (first vs. third person) AGAIN annotations are captured in an unbounded fashion which eliminates high degrees of reporting bias [7], [8].

The current dataset only encodes one affective dimension, arousal, across videos from 9 games; AGAIN, however, is easily scalable to more affective dimensions (e.g. valence or dominance) and more games-based affect stimuli. Future work will focus on expanding the labels with expert annotations of valence and dominance to match the format of other affective computing databases [59], [41], [43], [44]. Its accessibility and its unobtrusive data collection nature through crowdsourcing make AGAIN easily extendable to more affect labels, affect elicitors and participants.

Inspired by recent work on the importance of game context as a predictor of affect [3] the user modalities of AGAIN are currently limited to in-game video footage and behavioural telemetry data. In addition, the protocol of AGAIN limits the user modalities available so that first person crowdsourcing of affect annotations is both feasible and efficient. While AGAIN puts an emphasis on accessibility—soliciting game context and behavioural data from users as its modalities—the AGAIN games can be used for small-scale, lab-based affect studies that incorporate more user modalities including visual and auditory player cues (e.g. [49]).

Given the characteristics of a unique set of diverse elicitors, a large participant count, first-person annotations and a large-scale video and game telemetry database, AGAIN couples important aspects of affective computing with core aspects of game user modelling thereby enabling research in the area of general affect modelling, in games and beyond.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 programme under grant agreement No 952002.

REFERENCES

[1] J. Togelius and G. N. Yannakakis, “General general game ai,” in 2016 IEEE Conference on Computational Intelligence and Games (CIG). IEEE, 2016, pp. 1–8.
[2] E. Camilleri, G. N. Yannakakis, and A. Liapis, “Towards general models of player affect,” in Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), 2017, pp. 333–339.
[3] K. Makantasis, A. Liapis, and G. N. Yannakakis, “From pixels to affect: a study on games and player experience,” in 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019, pp. 1–7.
[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., “Human-level control through deep reinforcement learning,” nature, vol. 518, no. 7540, pp. 529–533, 2015.
[5] D. Melhart, A. Liapis, and G. N. Yannakakis, “Pagan: Video affect annotation made easy,” in Proceedings of the 8th International Conference on Affective Computing & Intelligent Interaction (ACII), 2019.
[6] P. Lopes, G. N. Yannakakis, and A. Liapis, “Ranktrace: Relative and unbounded affect annotation,” in Proceedings of the Intl. Conference on Affective Computing and Intelligent Interaction. IEEE, 2017, pp. 158–163.
[7] G. N. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions: An emerging approach,” IEEE Transactions on Affective Computing, 2018.
