The Australian National University
2600 ACT | Canberra | Australia

School of Computing

College of Engineering and
Computer Science (CECS)

Exploring Human emotional experiences in playing Overcooked under different difficulty settings

— 24 pt Honours project (S1 2022)

A thesis submitted for the degree
Bachelor of Advanced Computing

By:
Tong Cai

Supervisors:
Dr. Priscilla Kan John
Ms. Xuanying Zhu
Dr. Paul Wong

November 2022

Declaration:

I declare that this work:

- upholds the principles of academic integrity, as defined in the University Academic Misconduct Rules;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the class summary and/or Wattle site;
- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

November, Tong Cai

Acknowledgements

First of all, I want to express my sincere thanks to my supervisors Dr.Priscilla Kan Jon, Dr.Paul Wong and Ms.Xuanying Zhu for their top-notch instructions and guidance. They are incredibly amiable and have assisted me during my honour research, it is such a pleasure to be supervised and work with them. Moreover, throughout the entire research process, Ms.Xuanying Zhu has been incredibly understanding and supportive, She kindly provided lots of helpful suggestions and assistance both in my studies and life, and she made me realize the magic of Human-Computer Interactions. I also want to express thanks to my peer Fangxu Zhao, we have lots of conversations during the research, and he really supported me a lot.

I also want to express my thanks to my parents Ganlin Cai and Yanping Zhang for their care and encouragement. They always encourage me when I feel down and depressed. Being their daughter is my greatest luck, I'm grateful for everything they did for me. Thanks again for their unselfish support and love. I would also thank my nearest friends Siyi Wu, Wanyi Li and Yiwen Fang, thanks for their accompanies from a very young age and their warmest hugs are always waiting for me when I feel frustrated.

Abstract

Due to the rapid advancement of information technology, Artificial Intelligence(AI) is being utilised frequently in our daily lives. This research focuses mostly on Overcooked AI, a condensed form of a cooperative game that uses AI agents. Based on that, we investigated how various game difficulties impact players' emotional experiences and how to evaluate user emotional experience by the collected data.

To achieve this, we conducted a user experiment by inviting a variety of participants to play Overcooked with pre-trained AI agents under various difficulty conditions, where the game difficulties were changed by altering the game's layout and time duration. Furthermore, we measured the emotional experience of participants using their physiological data collected from the Empatica E4 wristband, as well as Self-Assessment Manikin(SAM) and online surveys during multiple game-playing sessions. From the results, participants' physiological data provided a trustful path for recognizing their emotions of valence and arousal states from SAM, an accuracy of 0.626 in assessing human emotions was obtained by Random Forest using. Moreover, the results exhibited that participants played Overcooked with AI agents in different difficulty conditions had emotions of different valence and arousal, where they felt the highest valence and arousal in medium difficulty condition. However, our evaluation only showed that participants played Overcooked with same settings in the second attempt had various levels of engagement and cooperation levels with AI agent, but didn't express a different emotion(no obvious changes in both valence and arousal), so it was seldom for participants to improve their game skills only after the first attempt.

Nevertheless, we concluded that physiological signals offer an effective method for evaluating user emotional experience. However, we still need to change Overcooked's current design and adjusting game difficulties to enhance user emotional experiences while playing cooperative games similar to it, since changing difficulty circumstances in cooperative games would cause diverse emotional experiences for users and we hope that participants can maintain a high level emotional experience in game playing sessions.

Table of Contents

Chapter 1

Introduction

Collaboration games are popular among the world, however, with the fast development and mature application of artificial intelligence, most collaborative game developers have considered or achieved adding AI agents as a choice of teammates for their intended users. For example, when users are about to play games, they can easily choose to play with real human or AI agents. Under this situation, a concern about the current design of AI is becoming more and more serious because we need to ensure its advantages comparing to real human to further ensure its application market boarder. Thus, our study concentrated on a collaborative game called Overcooked due to its easy operations to explore more aspects of Human-AI interactions.

”Overcooked AI”, which is a simplified version of the popular game ”Overcooked” was used in our research. It provides three different AI agents, but we focused on Human-aware agents since they have been approved to have best performances than the other two AI agents when teaming with human models in the past work (Carroll et al., 2019), as well as presenting more robustness (Knott et al., 2021) and interaction flexibility (Nalepka et al., 2021) in game playing sessions. Given that AI agent in Overcooked was constant, inspired by a previous work performed on a single-player game Tetris (Chanel et al., 2011), we chose to change game difficulties in the collaborative game to observe how AI performs in cooperation with users in various game difficulty conditions. Additionally, game duration and layouts were chosen to be the control factors of game difficulties because participants were required to think of growing levels of game strategies as the game duration decrease or the layout complexity increase for the purpose of cooking and delivering onion soups as soon as possible in Overcooked.

It is known that user experience is one of the significant indicators to evaluate Human-AI interactions, but this concept is somehow too vague because there are various of mechanisms to evaluate user experience. To avoid that, we mainly explored human emotional experiences by introducing two psychological properties **”Valence”** and **”Arousal”** to describe human emotions through rating the polarity and intensity of their feelings during game playing sessions. Comparing to the traditional emotion-rating methods using 6 basic emotion categories, this mechanism avoided errors caused by different cultural and social backgrounds of users (Herbon et al., 2005). Besides, participants were also required to wear a watch-like sensor called the Empatica E4 wristband to collect their physiological signals including Electrodermal Activity(EDA), Blood Volume Pulse(BVP) and Heart Temperature as another indicators to evaluate human emotional experiences in the game playing period with AI agents since the validity of such physiological signals in emotion recognition has been tested in the previous research (Jerritta et al., 2011).

Considering the emerging ethics issues, the user experiment in our research was held by inviting 30 participants (mean age: 21.5; gender: 13 females, 15 males, 2 others;) from university students only after the researcher's ethical application (with Protocol Number 2022/309) got approval from the ANU Human Ethics Office. In the user experiment, each participant was required to rate distinct difficulty levels (easy, medium and hard) for the three Overcooked game playing sessions in different game settings of duration and layout, and repeated playing these three Overcooked game sessions after a short break. Based on this, there were three types of data been collected during the experiment for each participant:

- **User emotions:** collected through Self-Assessment Manikin
- **Physiological data:** collected through the Empatica E4 wristband
- **Survey answers:** collected through the online survey on Qualtrics about difficulty ratings, engagement and satisfaction

After some basic data pre-processing steps for the collected data listed above, the data evaluation and analysis were expected to observe and check:

- **Physiological data in emotion recognition:** Although some previous work had made it convincing, our research still needed to prove its validity again because it was based on a new collaborative game and also had some differences from previous work on the provided categories of basic emotions and physiological signals.
- **Relations between user emotions and game difficulties:** Our research was not only motivated by exploring Human-AI interactions in collaborative games such as Overcooked, but also keened to introduce the importance of adjusting game difficulties due to the reason that we were investigating how user emotional experiences changes with the changes of game difficulties. Consequently, it was split into two sub research questions for better evaluation. Furthermore, we assumed that user skills of participants increase on the second game attempt in user experiment because they were expected to become more familiar with the game after the first game attempt. The hypothesis for each research question were shown as below:
  - **H1:** Participants play Overcooked with AI in various difficulty conditions

will have various emotions.

- **H2:** Participants are easy to feel different emotions in the second game attempt.

Our evaluation focused on two main categories, which were survey analysis and physiological data analysis respectively. Most gathering survey answers were subjective and we found that most participants were used to choose the median selection even we have told them try to avoid it beforehand. However, from analysis, they were still evident to show that participants play Overcooked in different game settings have different levels of engagement, satisfaction and cooperation levels. Moreover, we found that in the second game attempt, most participants presented a higher game skill levels as we expected, which resulted in various levels of engagement and cooperation compared to the first attempt.

The analysis on physiological signals was mainly performed to investigate human emotional experiences from their physiological data collected from the Empatica E4 wristband. However, although emotion recognition is a mature technique nowadays, the performances of our classification models (MLP, SVM, random forest and logistic regression) were not good as we expected when recognizing human emotions from their physiological data, we analyzed it was caused due to the limitations of the sensors and the bias caused by the median ratings of emotions in the survey. This unexpected results might also caused by the limited size of data-sets. Hopefully in the future work, we will invite more participants to engage in the study and also consider more model optimization methods to get better results. Additionally, we performed some statistical methods to verify our research questions both from the comparisons between physiological features extracted from game sessions with various difficulty settings and different game attempts, and found that participants had emotions of different valence and arousal when playing Overcooked in different conditions, but there was no obvious changes happened to their emotions in both game attempts.

Chapter 2

Background

2.1 Overcooked Game

- **Brief Introduction:**

  Overcooked is a popular cooking simulation video game available for collaborative patterns. Players are playing the role as chefs in the kitchen to cook, serve, and clean up when preparing meals in order to finish as many dishes as they can in the allotted time. However, collaboration is sometimes required for players to complete the meal of each order by deadline. Each order that is completed correctly gets coins, with bonuses for speed, whilst orders that are incorrectly completed do not, and just serve to waste time. The objective in playing Overcooked is to gather as much coins as you can in the allotted period.

  There are various of kitchen layouts listed in Overcooked with different positions of each object like cooking pots and serving counters. However, the stations for ingredients, preparation areas, stoves and ovens, serving windows, and dishes are spread out over the kitchen, which requires participants to move between them and the movement often takes time. Moreover, there are also other challenges in each kitchen layout of Overcooked, such as trunks on the transmission road.

- **Human-AI collaboration:**

  Different from the Overcooked game on Streams, our research adopted a simplified version of Overcooked currently only supports Human-AI playing mode, which forces human players to collaborate with AI agents to cook and deliver meals in the restricted time duration. Hence, we want to explore more aspects about human-AI collaboration efficiency. More specifically, both human players and AI agents must be able to recognise one another's activities, anticipate the next move, and take matching measures in order to efficiently cook and deliver meals by the deadline.

However, testing Human-AI cooperation efficiency requires a more complex game environment because they can easily achieve the cooperation goal in a simple environment without any flexibility (which becomes more serious when we want to test AI's performance in adapting human behaviours).

Overcooked is helpful for us to enhance our studies in Human-AI collaboration due to its cooperative nature of the game's purpose (i.e., identify and predict the movements of partner and make adjustments) and the context's explicit definition of success (i.e., cook and deliver soups as much as possible in the given time length).

2.2 Ethics in data collection

In this research, it is crucial for us to conduct a user experiment to gather information from participants in order to more thoroughly analyse and assess human-AI interactions in Overcooked. But with human-centered computing, taking data collecting ethics into account is both a difficult necessity and a crucial factor. Moreover, ethics is not only considered in the field of HCI, but also applied to many areas which involves data science. We must focus on data science ethics beyond a review of these potential problems since data science activities threaten our understanding of what it means to be a person. As a result, the hazards of data science without consideration for ethics are becoming clear.

- **Data ethics:**

  If all data collected by human are made public on the internet, especially those protected data like names and addresses, then there might cause harm to humans whose information is public. Besides, it will also results in data leaking. To handle such data security problems caused by the lacking of ethics, embracing safe computing practises, carrying out regular system audits and adopting policies are becoming more and more important.

  Privacy, bias, access, personally identifiable information, encryption and legal requirementsm as well as constraints (spa, 2021), and potential problems are all ethical use factors. Furthermore, data ethics also entails being frank about the risks and repercussions that the data's subject matter and the organisations that use it may face (spa, 2021).

  In our research, to protect human data to the greatest extent, we submitted the ethical application to the ANU Human Ethics Office with mentioning our operations on collecting and protecting human data. For example, participants are not required to provide any identifying information during the experiment and the collected data won't be made public, it will only be used for research and protected safely.

- **Principles of data ethics:** (spa, 2021)

  1. **Transparency:** involves letting participants know what data is being collected

as well as what data is being used in our research.

2. Fairness: check biases in the rules that are applied, the questions that are asked of the data, and the data that is collected.

3. Accuracy: depends on a number of factors, including whether the data is accurate and if it makes sense and is helpful in light of what we are attempting to accomplish or learn.

4. Privacy: depends on a variety of variables, such as whether the data is reliable, if it makes sense, and if it is useful in light of what we are trying to learn or do.

5. Accountability: establishing policies to ensure that new tools and technologies are produced ethically, identifying and analysing risks and potential repercussions of producing and using data and AI, and considering the reliability and quality of data sources are all necessary.

2.3 Valence and arousal of emotion

There are various of approaches to evaluate human emotions in psychology, valence and arousal are two effective measurements among them and will be used in this research to explore human emotion experiences when playing Overcooked with AI agents. In the start, being familiar with the basic concepts of valence and arousal is helpful for the further emotion measurement.

- **Valence** In psychology, valence is an emotive property that describes the inherent appeal or ”goodness” (positive valence) or averseness or ”badness” (negative valence) of a situation, an item, or an event (Frijda et al., 1986), which also describes and classify specific emotions. Generally, emotions can often be split into two categories: positive and negative emotions. In this case, the positive emotions like ”enjoyment” and ”excitement” are usually seen to have positive valence while those negative emotions like ”anger” and ”sadness” have negative valence. Thus, valence-based approach is widely-used because emotions with similar valences are seen to have similar effects on judgments and choices when studied in relation to affect, judgement, and decision.

- **Arousal** Arousal indicates a condition of awakening or stimulation of the senses to the point of perception physiologically and psychologically, as well as involving activating the ascending reticular activating system (ARAS), which controls the endocrine, autonomic, and wakefulness systems in the brain. This causes a condition of enhanced sensory acuity, desire, mobility, and response readiness in addition to increased heart rate and blood pressure (Frijda et al., 1986). Generally speaking, emotions could be classified into two categories: emotions with high intensity and low intensity, and arousal is effective to describe the intensity of emotions. For example, the two emotion nouns ”a bit happy” and ”very happy”

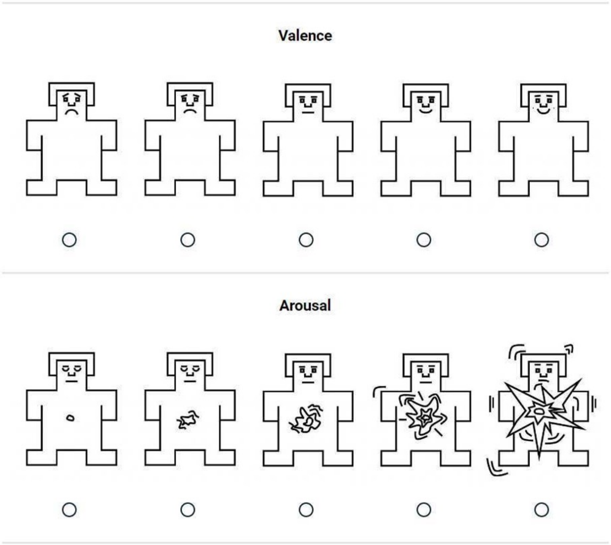

Figure 2.1: SAM used in this research (Figure from (Stevens et al., 2017))

present emotions in different intensity scales, and we could say that ”a bit happy” is the emotion with low arousal while ”very happy” is the one with high arousal.

However, in our research, to improve the precision in defining human emotions, we will combine valence and arousal as opposed to measuring emotions alone by valence or arousal. To be more detailed, we will use Self-Assessment Manikin to access human self scales in valence and arousal, where Self-Assessment Manikin is a non-verbal graphical assessment tool that quantifies a person’s affective response to a variety of stimuli by measuring their level of pleasure and arousal (Bradley and Lang, 1994).

And then match their emotions in the two dimensional valence-arousal space based on the valence and arousal ratings.

2.4 Physiological signals for emotion recognition

Emotion recognition based on physiological signals has been applied on many researches in the field of Human-Computer Interactions, we also adopted the approach to evaluate human emotional experiences in cooperation with AI agents. Comparing to other mechanisms such as gestures and facial expressions which are also valid in recognizing human emotions, changes in physiological rhythm is unavoidable and also easier to be detected because the sympathetic nerves of the Autonomous Nervous System(ANS) are activated when a person is positively or adversely aroused, and such sympathetic activation boosts blood pressure, respiratory rate, heart rate, and blood pressure variability (Jerritta et al., 2011). In our research, we mainly concentrate on measures of psychophysiology such as: 1. Cardiovascular system: Heart Rate(HR), Blood Volume Pulse(BVP); 2. Electrodermal Activity(EDA); 3. Skin Temperature.

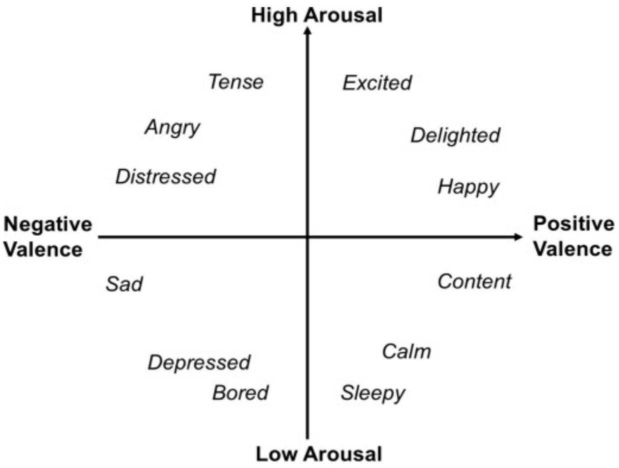

Figure 2.2: two dimensional valence-arousal space (Figure from (Du et al., 2020))

However, although different emotions have relationship with the changes of the listed physiological signals, features extractions is necessary in emotion recognition model, which forces us to to identify the most emotional-relevant characteristics and link them to emotional states. The mostly-used features extracted from physiological signals are: mean, standard deviation, first derivative, high frequency and low frequency powers, Fourier and Wavelet Transform coefficients etc (Wioleta, 2013).

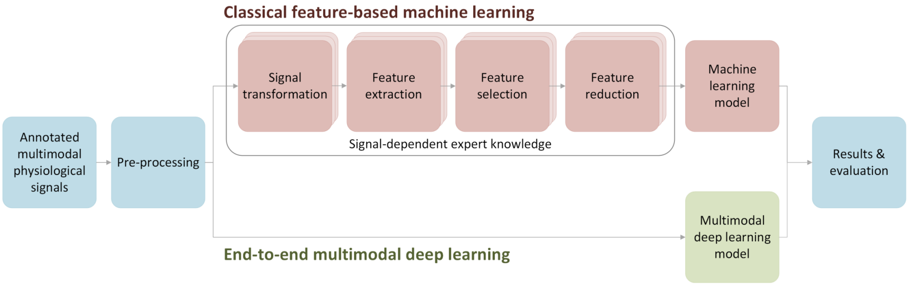
Classical feature-based machine learning
Signal transformation
Feature extraction
Feature selection
Feature reduction
Machine learning model
Signal-dependent expert knowledge
Annotated multimodal physiological signals
Pre-processing
Results & evaluation
Multimodal deep learning model
End-to-end multimodal deep learning

Figure 2.3: example of emotion recognition using physiological signals (Figure from (Dzieżyc et al., 2020))

Also, from the previous researches, it showed that neural network, random forest, SVM were the three common-used classification models to solve emotion classification problems (Zhao et al., 2018). And the performances of emotion recognition using physiological data replied on multiple factors, for example, the extracted features from the physiological data, the selected features and labels for solving classification models might had influences (Zhao et al., 2018).

Chapter 3

Related Work

3.1 Studies of AI implementations

Since Overcooked is a conventional cooperative game, it has been the subject of numerous studies, the most of which have focused on testing and analysing the AI implementations, which are an important part of Overcooked's design.

- **Human models**

  Human errors greatly reduce the effectiveness of maximizer agents at the team level in cooperative environments, and policies established by agents may not be clear to human players, which leads to failure if the human doesn't perform their expected role in the cooperation.

  In this study(Carroll et al., 2019) , various kinds of AI agents are originally introduced alongside human models (agent designed for self-play and real humans, those agents are developed by different planning algorithms). The cross-entropy loss on data involving human teams and various types of AI was then seen through user research. Additionally, a comparison of various agent collocations was done using the average payout each episode. The findings make it clear that artificial agents that have learned from humans exhibit greater adaptability and have the capacity to behave as both leaders and followers than agents that expected their companions to perform more capably.

- **"Interaction flexibility of AI in teaming with humans**

  (Nalepka et al., 2021) was known about how to create these adaptable and flexible coordination patterns, so the experiment described in the paper asked 43 undergraduate students to play with two different kinds of artificial agents in Overcooked—the self-play agent and the human-aware agent, respectively—in order to discuss the significance of creating artificial agents that can support and improve team coordination.

The assessment is based on the game score, the flexibility of the interactions (%Determinism), and a qualitative measure of interactions (five-point Likert scale questions rated by participants). According to the research, playing with human-aware agents tends to increase the likelihood that game tasks will be completed successfully. In addition, human-aware agents perform better in task completion and teaming experiences and exhibit greater flexibility than self-playing agents. Another finding in the results is that for both artificial entities, the quantity of offered onion soup bowls is inversely correlated with their interaction flexibility. Furthermore, interactions between participants and human-aware entities show more flexibility when focusing just on the transition states within teams. In other words, individuals typically perform better and engage in more flexible interactions when teaming with human-aware AI.

- **robustness of collaborative agents**

  Due to the variety of the real world and the complexity of human behaviour, robustness has become an important characteristic of collaborative agents in the game. Therefore, it is crucial to take into account as many edge cases as possible in the Overcook game in the test suites in order to properly assess the robustness of the agents.

  The research employs an experiment to assess three ideas to enhance robustness Knott et al. (2021) : ToM-enhanced human model quality, increased model diversity, and state diversity (starting from states visited in human-human game play), respectively. The average reward of the agents when paired with human models from a suite of 20 validation agents is computed in this experiment using four different Overcooked layouts, 50 roll-outs on the unit tests, and four different Overcooked layouts. The unit test robustness, based on the results, exhibits various corresponding relationships with the average reward in these approaches, but the population of 10BC and 10ToM agents produces the best results.

  The research also addressed the difficulty of evaluating the robustness of collaborative agents in situations when the proper behaviour is uncertain; in order to improve robustness, we may need to take into account more potential behaviours in unit tests and different game layouts.

- **"Diversification Methods in Cooperative Multi-agent Deep Reinforcement Learning"**

  The experimental findings are used in this research to compare and analyse various diversification strategies. The primary driving factor behind this is that co-evolution of learning agents causes Over-fitting in multi-agent reinforcement learning, necessitating the introduction of diversity. However, commonly used approaches are unreliable when doing so under fully cooperative multi-agent MDP.

The paper Charakorn et al. (2020)proposes four distinct methods: self-play, partner sampling, population-based training (PBT), and pre-trained partners. The average cumulative payout of self-play (teams of the same agent) and cross-play (teams of the same type but different agents) teams of agents is compared in an experiment utilising each of the four unique agents with unidentified partners in the Overcooked scenario. The results show that self-play and partner sampling agents, as well as PBT agents, are unreliable sources of diversity in cooperative games because they only function well with agents from the same population. Pre-trained agents with distinct hyper-parameters outperform other agents and score higher in cross-play when ad hoc teaming is present. The possibility that the collaborative game will promote variety will thus be increased by the use of pre-trained agents.

3.2 Evaluating user experiences in games

While gathering user game experiences is important, user experiments are frequently employed. However, there are many different ways to gather and explore user experiences, such as physiological signals, questionnaires, and emotion evaluations. In this area, numerous studies have been conducted.

- **effective approaches to structure and represent emotions**

  Herbon et al. (2005)In the past, participants were typically asked to describe their feelings using emotional words like "happy" or "angry" on their own in order to quantify their emotions in a particular field. Labeling with emotive terms is not very accurate when examining human emotions, though, as many people may experience a range of feelings. In addition, participants may have diverse perspectives on fundamental emotions due to differences in their cultural or social backgrounds, which necessitates the usage of various fundamental emotion categories for various research subjects.

  To improve the accuracy of assessing human emotional experiences, a new method of exploring valence, arousal, and dominance is introduced in this case. Valence refers to the positive or negative aspects of an emotion, whereas arousal can range from excitement to relaxation, and dominance can range from feeling submissive to having control. A user experiment was carried out by allowing students of various ages and genders to rate their valence, arousal levels in SAM(Self-Assessment Manikin) while playing games with increasing levels of difficulty in order to further investigate the applicability of this approach in assessing human emotions. Although there are still some problems with inter-individual differences and ensuring the participants' honesty during self-evaluation, the approach is still in some ways valid. This work, however, only focuses on two-dimensional (valence-arousal) emotion models; it is still uncertain whether there are higher dimensional models that may assess user emotions by adding a new dimension like dominance.

- **Exploring eye activities**

The study of eye activity of human could be one persuasive mechanism to evaluate user game experiences. The primary focus of this work Chen et al. (2011) is on the effectiveness of pupillary responses to assess player mental effort in games, which roughly includes blink latency, blink rate, average pupil size before and after, standard deviation of pupil size, fixation time, fixation rate, saccade size, and saccade speed.

In this study, a user experiment is carried out by asking participants to recognise defenders and attackers in a video at progressively higher levels of cognition and to recollect their locations at the conclusion of each 15-second clip.

According to the experiment's findings, participants' eye movements changed when switching from a low-level task to a high-level task: their blink latency, pupil size, and fixation duration increased, while their blink rate, pupil size deviation, fixation rate, saccade speed, and saccade size decreased. This confirms the usefulness of blink and pupil response as workload indicators.

However, limits still exist because each participant's feature value ranges are different, hence in this experiment, not all characteristics benefit from calibration. In this instance, calibration enhancement of measurement is significant.

- **Emotion assessment from physiological signals**

  This work Chanel et al. (2011)primarily focuses on the physiological signals from the peripheral nervous system and central nervous system. Physiological signals are frequently employed in analysing user game experiences. Additionally, this work suggests some emotional states utilising two-dimensional valence-arousal scales, such as boredom, engagement, and anxiety.

  In this research, a user experiment was conducted to examine how people interacted with Tetris games at various levels of difficulty, where the speed of the falling blocks determines how challenging the game is. To more thoroughly investigate how game difficulty affects player emotions and how to modify game difficulty to better match user skill growth, This study compared user emotions across a range of user skill levels and difficulty situations. The results demonstrate that individuals experienced decreased valence in medium conditions compared to easy and hard conditions, whereas arousal levels increased as difficulty increased. Additionally, it demonstrates that as user expertise increases, valence and arousal both drop.

  However, it still lacks accuracy in predicting users emotional states using their physiological signals.

The current simplified version of Overcooked features five different layout options and three pre-trained AI agents, two of which have learned through self-play and the third from real people. The related work mentioned above (Carroll et al., 2019) (Knott et al., 2021) (Nalepka et al., 2021) has discussed in evaluating various AI implementations and comparing their performance, either in terms of their ability to adapt to human behaviours or their flexibility in terms of interaction, etc. According to the published results of the related work, the AI agent that was taught by humans performed better in Overcooked than the two other types of AI that were taught by self-play.

In this research, we will only take into account testing human-aware AI agents while putting this concept into practise. Furthermore, in contrast to the state of the art, our present ideas are more concerned with how users interact with AI agents when playing games than with how well they do while interacting with actual people.

Our strategy focuses mostly on investigating how players react to playing Overcooked at various levels of difficulty and then further adjusting the level of difficulty to improve user experience. Then, as the two variables that will regulate game complexity, game layouts and time limits were selected.To accomplish the objective of providing soups as quickly as possible, the user experience would be negatively impacted by a poorly planned layout since users would need to constantly adjust to it, which would increase the game's difficulty. In Overcooked, shorter time lengths may cause the difficulty to increase while longer time periods cause it to decrease. Our current plans call for creating three game difficulty levels (easy, medium, and hard) based on the various ways that layouts and game duration might be combined. The assessment is based on the users' physiological signals that have been gathered and their questionnaire responses. The self-emotional assessment that participants perform using SAM is a crucial measurement for acquiring participants' game experiences.

Chapter 4

Experiment Design

4.1 User experiment design

The experiment was held based on a simplified version of Overcooked game, since we were interested in exploring Human-AI interactions in cooperative games and Overcooked features straightforward testing environments, it would be helpful to limit the impact of any other aspects in our research.

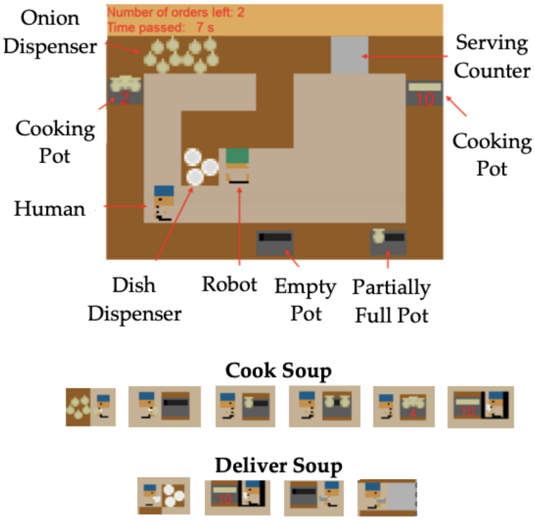

Figure 4.1: basic Overcooked environment, (Figure from (Fontaine et al., 2021))

Additionally, Overcooked was one of the most played games worldwide, and most gamer

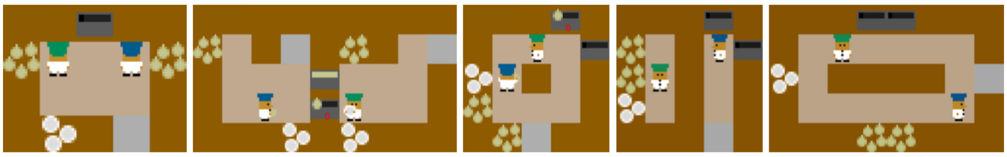

Figure 4.2: layouts provided by Overcooked. (Figure from (Carroll et al., 2019))

have played it before with a partner, so they were familiar with the game's features. However, the actions required to play Overcooked were also straightforward. Participants simply needed to press the "left," "right," "up," and "down" keys on their laptops to control the movements of their playing character. Also, they can directly press the space-bar when facing objects (onions, plates or onion soups) and interacting with them.

- **Factors to control game difficulties**

  We offered three distinct time lengths to distinguish game complexity: 30 seconds, 45 seconds, and 60 seconds. Since layouts and time lengths were the two main criteria to determine game difficulty, recalling there were five alternative layouts in Overcooked, but testing all five might left people feeling bored and worn out, so we only tested three of the five layouts in the experiments, which are Cramped Room, Coordination Ring, and Counter Circuit respectively.

  Control Factor-1: Game layouts: According to our assumptions, various game layouts resulted in various game strategies. Since Overcooked was a cooperative game, players can anticipate to have varying degrees of collaboration with AI agents in various game layouts, which will, in part, affect how tough they found each layout to play. In addition to the three layouts we selected above, Overcooked also offered the layouts Asymmetric Advantages and Forced Coordination.

  While the three chosen layouts present participants with escalating collaboration level obstacles, these two layouts didn't reveal any appreciable cooperation level discrepancies with others. Moreover, the published results Carroll et al. (2019) indicated that the performances of most models are significantly low in Counter Circuit.

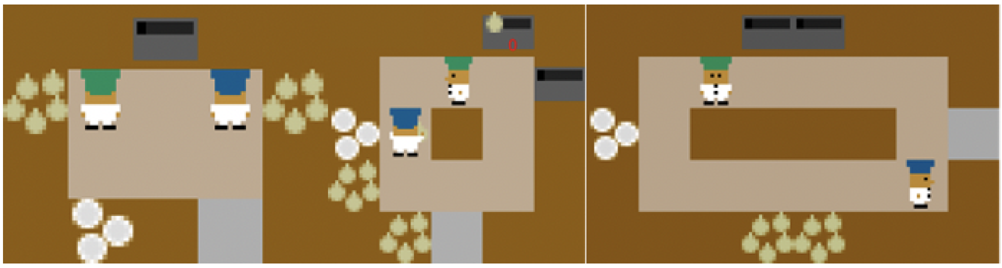

Figure 4.3: the three layouts chosen in the experiment, (Figure from (Carroll et al., 2019))

- **Cramped Room:** This layout only provided one cooking pot and the serving counter locates at the opposite edges of the room, the onions were placed on both sides of the room and plates were available next to the serving counter. It presented low-level collaboration levels (Carroll et al., 2019)between humans and AI agents because neither of them needed to cooperate with the other to achieve the goal of cooking and delivering onion soups. It was chosen to represent the easy game difficulty conditions and we expected that most participants rate it as the easy layout while keeping other variables consistent(like AI agent and time duration).

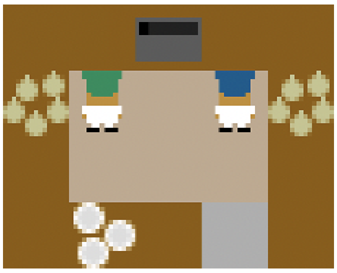

Figure 4.4: Cramped Room, (Figure from (Carroll et al., 2019))

- **Coordination Ring:** Different from Cramped Room, Coordination ring offered two cooking pots on both sides of the upper right corner of the room, onions were only placed on the left side of the room while plates were available on both sides, also, the room seemed more like a "ring" instead of a "rectangle", which forced participants to move between the bottom left and top right corners of the room (Carroll et al., 2019) for cooperation to cook and deliver the onion soups. We expected participants to have mid-level cooperation challenges when playing in this layout and rate it in medium game difficulty.

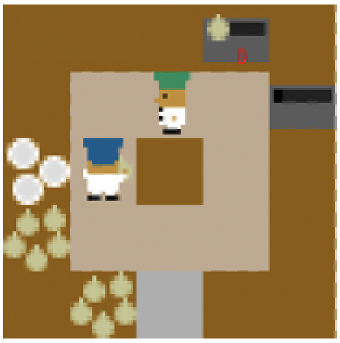

Figure 4.5: Coordination Ring, (Figure from (Carroll et al., 2019))

- **Counter Circuit:** This layout presented a larger "ring" than Coordination Ring, however, the two cooking pots were fully-connected on the upper side while the cooking counter was on the right side of the room. Since the onions were placed on the left side and the plates were on the lower side in the layout, it increased the difficulty for participants to think of the optimal game strategy to cooperate with AI agents because they were very easy to collide without the optimal cooperation strategies. We expected participants rate it as the most difficult layout.

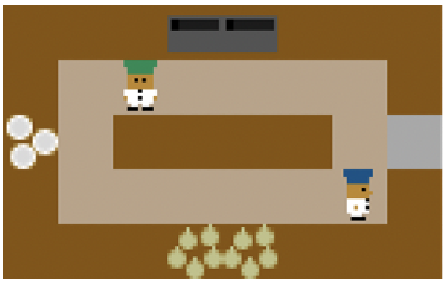

Figure 4.6: Counter Circuit, (Figure from (Carroll et al., 2019))

It was worth mentioning that the biggest difference between Cramped Room and the other two chosen layouts (Coordination Ring and Counter Circuit) was that those two layouts require higher cooperation levels between human and AI agents, in this case, human and AI agents were required to identify and predict each other's behaviours and make corresponding adjustments to reach the optimal cooperation strategy.

Control Factor-2: Time duration: Regarding our decision to choose time lengths, since participants must work with AI agents to quickly cook and deliver soups, longer time may make them boring because they were continuously completing the same actions without any variation in the game, while shorter time increased anxiety because they have a limited amount of time to play games. Before the research, we have tried various of time lengths, and found these three time duration (30 seconds, 45 seconds and 60 seconds) could successfully distinguish three different difficulties in Overcooked.

To test all layouts and time lengths, there were nine different combinations. Considering participants will easily get bored if testing all the combinations, so we only intended to invite each participant to test three combinations among all instead to reduce boredom.

- **Data Collection**

  We mainly collected participants' physiological data and their answers to the online survey questions.

  Physiological data: As discussed above, physiological signals could somehow be a valid measurement to evaluate user game experiences because they were difficult to fake. In the experiment, we will focus on the physiological signals from the peripheral and central nervous systems. To achieve this, participants will be asked to equip with the Empatica E4 wristband to detect: 1. Electrodermal activity (EDA) reflects how sweat the skin is; 2. The blood volume pulse (BVP) reflects the changes in the volume of blood, which is an effective way to measure heart rate; 3. Skin Temperatures; 4. Accelerometers capture motion-based activities.

Online surveys: We also offered an online survey on Qualtrics to collect more thorough data to analyze player game experiences. Participants will be first asked for basic background information in the survey, including their age, gender, area of study, and prior Overcooked game experiences. Participants will then be asked to rate the game's difficulty, level of AI collaboration, and level of participation for each session using 5-point Likert Scales. In addition, we asked participants to answer questions about their preferences for the given layouts and time limits as well as their suggestions for enhancing game experiences in order to measure our control of difficulties and to bring participants better game experiences after some potential future improvements.

Experiment Recruitment: Participants were recruited from social medias like Wechat, Facebook and ANU SONA systems.

However, there were some limitations for participants:

1. **Participants enjoy playing games and it would be better if they have experiences in playing collaborative games.** Since we aim to explore how participants' emotional experiences change in different difficulty conditions. If participants don't even enjoy playing the games we want them to play, they are less likely to exhibit happy emotions in different game-playing sessions, regardless of how the game difficulty changes, which increases the bias for our research. Moreover, participants will work with pre-trained Human-aware agents in Overcooked, if they have prior experience with collaborative games, they may be better used to the game's mechanics.

2. **Participants should have good health conditions.** Since collecting physiological data is a hard requirement in our research, if participants have any diseases such as cardiopathy or hypertension, the quality of collected physiological data (e.g. Blood Volume Pressure and Heart Rate) from them will be be affected.

3. **Those participants aged below 18 are excluded.** We must get approval from the guardians if children (aged below) anticipate our experiment, which aggravates the ethical considerations.

Additionally, we will try our best to reduce the gap in the number of male and female participants to avoid gender bias.

Emotion assessment: As we provided three difficulty conditions (Easy, Medium and Hard) in Overcooked, by observing the changes in participants' physiological data under various difficulty conditions, we can then speculate how participants emotional experience changes according to that. Although physiological data is hard to fake by participants, we could still use some other approaches to acquire participants emotions. According to the related work Herbon et al. (2005), we will ask participants to rate their feelings on Self-Assessment Manikin (SAM). In our research, we will only considered the two-dimensional valence-arousal space instead of the three-dimensional valence-arousal-dominance space, because adding a new dimension might increase the complexity while the two dimensions are enough to present emotions precisely.

In this case, we currently constructed three emotion categories based on different valence and arousal ratings in the two-dimensional space: 1. Boredom (negative valence and low arousal); 2. Engagement (positive valence and high arousal); 3. Anxiety (negative valence and high arousal). At this stage, because it was less often than the other three emotion categories when we examined players' emotions while playing games, we didn't take the emotion category in positive valence and low arousal like calm into consideration.

Experiment process: After getting approval from the ANU Human Ethics Office, we recruit 30 participants(mean age: 21.5; gender: 13 females, 15 males, 2 others;) from university students to participate in the study. At the beginning the experiment, participants were required to read the participant information sheet to get familiar with what involves in this study first and then signed the written consent form to confirm their agreement to take part in the research, as well as ensuring each participant was in good health condition and also have normal or corrected-to-normal visions.

Following that, participants must adhere to instructions and put the Empatica E4 wristband on their hands and then began to play Overcooked on the online demo available at: Overcooked-demo. The whole experiment contained 6 sessions of Overcooked game, three of which were performed using different combinations of game layouts and time duration and the other three of which were repetitions of the first three to further examine whether user skills would improve if they played the same game twice (which might potentially lead participants to change from engagement to boredom). In this case, we set a rest time of approximately 1 minute to allow participants to rate their emotions, difficulty level, etc. in the survey and also transmit the physiological data collected in each session to the laptop during this time in order to get more accurate physiological data from the wristband and also leave enough time for participants to rest and fill out the corresponding parts of the questionnaire, the last but most important reason to set the short break was to let participants' physiological signals return to the normal state.

Moreover, each session last for no longer than 2 minutes, and the researchers were responsible to change the set-up components (combinations of game layouts and time lengths) in Overcooked based on the orders of combinations for each participant to play.

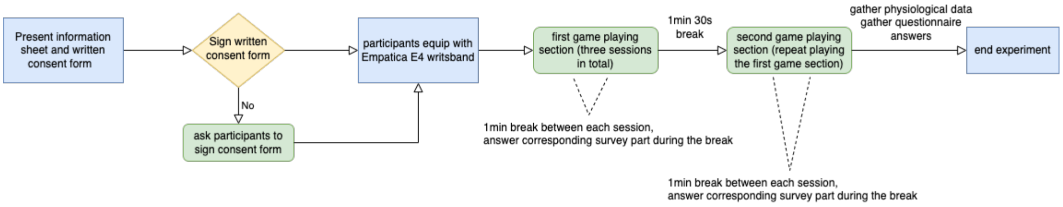
Present information sheet and written consent form
Sign written consent form
No
ask participants to sign consent form
participants equip with Empatica E4 writsband
first game playing section (three sessions in total)
1min break between each session, answer corresponding survey part during the break
1min 30s break
second game playing section (repeat playing the first game section)
1min break between each session, answer corresponding survey part during the break
gather physiological data gather questionnaire answers
end experiment

Figure 4.7: Experiment flow chart

4.2 Research questions

Participants' physiological data was collected, and their responses to the questionnaire allowed us to determine their emotional states. We need to categorise physiological data into three separate emotion categories—boredom, anxiety, and engagement—in order to confirm the viability of physiological signals in distinguishing human emotions. We trained various classification models to achieve it utilising the retrieved physiological data characteristics as inputs and user emotional states as goals from each Overcook game session in the experiment. Also, we need to determine human emotions with the changes of game difficulties,

Based on that, we proposed three research questions and corresponding hypothesis:

- **Research Question-1:** Participants' emotional states could be assessed from the physiological data.
- **Research Question-2:** Will participants present different emotions in different game difficulty conditions?

  Hypothesis: Participants present different emotions when playing Overcooked in difficulty conditions.

  In our experiment design, each participant was randomly assigned to play three sessions of Overcooked in different combinations, and after playing, they had to rate the difficulty level for each game session. To avoid participants giving the same difficulty ratings for multiple game sessions, we will inform participants to try to distinguish different difficulties in each session as much as possible even if their subjective differences between each of them were quite tiny.

  Moreover, not only participants' emotions were collected during the break between each session, but also their physiological data for the corresponding session was collected. Nowadays, knowing physiological data was efficient in emotion recognition, so to answer this research question and verify our hypothesis, we need to analyze and compare their emotions.

- **Research Question-3:** Comparing to the first game attempt, will participants find themselves have different emotions in the second game attempt?

  Hypothesis: participants will switch from the original emotions into another one in the second game attempt.

  We assumed that participants' user skill would increase when they try the game in the same difficulty conditions multiple times. To verify this assumption, in the user experiment design, each participants were required to play these three sessions again after a 1min30s break. After that, we can compare their emotional states from both game attempt to conclude whether there is an obvious difference.

  Furthermore, similar as what the experiment design covered for the first research question, participants' physiological data were also collected for each session in both game attempts, which was somehow a valid approach in determining whether participants feel boring in the second game attempt. For example, when participants get bored, their physiological data such as Heart Rate and Skin Temperature might decrease from the state where they are engaged.

Chapter 5

Methodology

5.1 Data Pre-processing and Feature extractions

Data pre-processing After gathering data of all participants in our user experiment, to keep the data quality and ensure its validity in evaluation, we have to take some pre-processing steps on the raw data ahead.

- **Remove noise:**

  Taking the raw physiological data collected from one participant in one of the game sessions as an example, from the boxplots showing below, we can observe that there exists outliers in the raw EDA, BVP and Skin Temperature data.

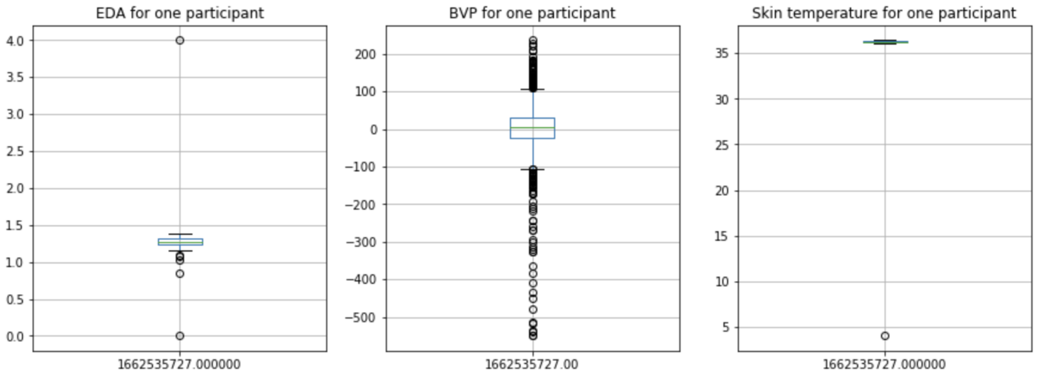

Figure 5.1: Boxplots for raw physiological data of one participant

  To remove these noise in the raw data and achieve data cleaning, 3-sigma rule is adopted to deal with them. Since the collected physiological data should obey the standard normal distribution, and most of them should distribute in the range $(\mu-3\sigma, \mu+3\sigma)$ based on the standard normal distribution, then 3-sigma rule defines a value in the raw data not within this range as abnormal value and remove it where μ and σ are calculated based on our raw data. We implemented the 3-sigma algorithm on each category of raw physiological data to remove noise.

- **Data encoding:**

  Most data we collected is numerical, but strings represent the participants' emotional states as determined by their judgments of valence and arousal as well as the degree of difficulty for each game session. Moreover, our research is about to find the relations between participants' physiological data, emotional states and game difficulty throughout various game playing sessions. In this case, due to the reason that strings takes up more memory than integers, we encode the string objects into numerical one based on the table below to increase efficiency.

  Table 5.1: Data encoding table

Game difficulty	Emotional state	numerical number
Easy	Boredom	0
Medium	Engagement	1
Hard	Anxiety	2

- **Normalization:**

  We normalize all numerical data (except game difficulties and emotional states of participants) into values range between 0 and 1 to unify statistical distribution of induction samples.This is necessary because some of the features of our data may have different value ranges, which requires us to make them in the same measurement scales.

Physiological feature extractions

- **Electrodermal activity(EDA):**

  The collected human EDA responses can be splited into two categories:

  1. **tonic change:** describes the slow and smooth changes in the EDA response signal that take place in the absence of stressful stimuli (Airij et al., 2020).

  2. **phasic change:** describes fast/quick variations in the EDA response (Airij et al., 2020).

  However, since our research focuses on exploring human emotional experiences during Overcooked game playing sessions, then analyzing phasic changes of EDA responses would be more valuable in recognizing human emotions because it indicates fast changes in EDA while tonic changes are relatively slow.

Thus, inspired by the thoughts, our ideas of extracting peaks-related features came from pyEDA, which is an Open-Source Python Toolkit that supports feature extractions from EDA (Aqajari et al., 2021). Moreover, pyEDA is able to pre-process the raw EDA data and get filtered EDA, so we won't use 3-sigma mechanisms to deal with outliers for EDA. Given that the collected EDA data from Empatica E4 wristband has the sample rate at 4HZ, and we defines the segment width as the length of each EDA data because we don't consider the segment operations at this stage. The extracted features from EDA are listed in the table below.

Table 5.2: extracted features from EDA

extracted EDA features	
Feature-1	number of peaks
Feature-2	mean eda value
Feature-3	max amplitude of peaks
Feature-4	mean filtered phasic eda
Feature-5	mean phasic eda
Feature-6	mean tonic eda
Feature-7	mean peak list
Feature-8	mean indexlist

- **Blood Volume Pulse(BVP):**

  BVP is sensitive to deviate Heart Rate Variable(HRV), thus, to enhance its ability in emotion recognition, we mainly focused on those features related to HRV instead of statistical features extracted from other physiological data like mean value, standard deviation etc.

  The features extraction from BVP was splited into two steps:

  1. **Get clean BVP:** The quality of the BVP data will also affects the quality of extracted features. For this reason, inspired by pyHRV (Gomes et al., 2019), which is a popular tool for preprocessing heart related data, we used the modules inside biosppy.signals to get the filtered BVP data. Compared to 3-sigma method, this module is more helpful to make BVP data more smooth on the basis of outliers removal. Taken one series of BVP data collected from one participant in a game-playing session as an example, the raw BVP data and the filtered BVP data are shown as the figures below.

  2. **Feature extractions:** The ideas of BVP related features extractions comes from HeartPy, which is a Python Heart Rate Analysis Toolkit module in Heart Rate Analysis (Van Gent et al., 2019) (van Gent et al., 2018). The collected BVP data from Empatica E4 wristband has a fixed sampling rate of 64 Hz, through using the package inside HeartPy, we acquired both working data (like peaks positions

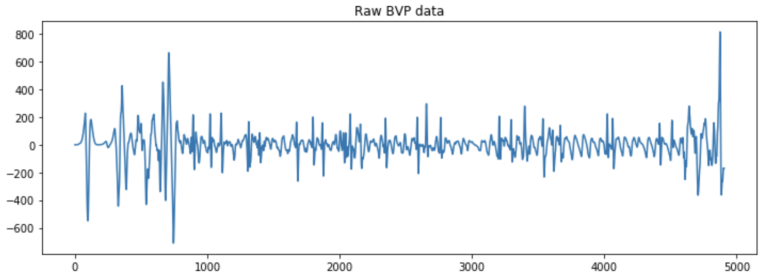

Figure 5.2: Raw BVP data.

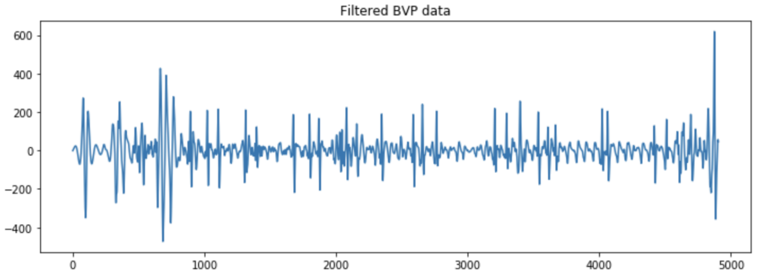

Figure 5.3: Filtered BVP data.

and intervals between peaks) and computed output measures from the filtered BVP data on the last step.

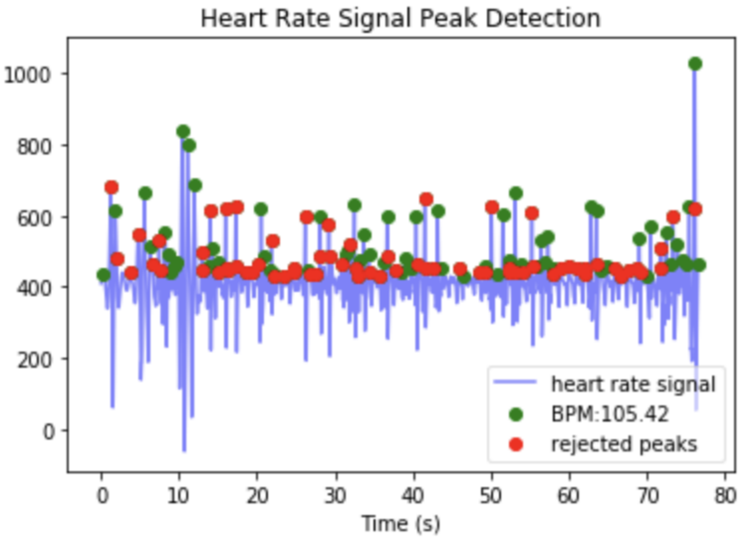

Figure 5.4: BVP peak detection

The output time-domain parameters are listed in the table below:

Table 5.3: extracted features from BVP

	extracted BVP features
Feature-1	proportion of differences between R-R intervals greater than 20ms (PNN20)
Feature-2	proportion of differences between R-R intervals greater than 50ms (PNN50)
Feature-3	beats per minute (BPM)
Feature-4	interbeat interval (IBI)
Feature-5	standard deviation if intervals between adjacent beats, SDNN
Feature-6	standard deviation of successive differences between adjacent R-R intervals (SDSD)
Feature-7	root mean square of successive differences between adjacend R-R intervals (RMSSD)
Feature-8	median absolute deviation (MAD)
Feature-9	Poincare analysis (SD1, SD2, S, SD1/SD2)

- **Skin Temperature:**

  We calculate the following features for skin temperature (Chanel et al., 2011):

	extracted features from SKT
Feature-1	mean value
Feature-2	standard deviation
Feature-3	mean standard deviation

5.2 Emotion recognition

Our research aims to verify the validity and efficiency of physiological data in recognizing human emotions when participants are playing Overcooked with human-aware AI agents in different game difficulty conditions. Based on that, we devised a multiple classification problem by giving the extracted physiological features as inputs and emotion categories as label outputs. To achieve it, we trained several classification models which are popular nowadays to solve such multi-classification problems and adopted different evaluation metrics to compare their performances more comprehensively.

Besides, since distinct physiological signals might present various performance in emotion recognition, we also designed different set of the extracted physiological features as various inputs to further investigate it.

Inputs and Label Outputs:

The physiological features are mainly extracted from the three kinds of physiological data, which are BVP, EDA and Skin Temperature respectively. However, since Skin Temperature are very likely to be influenced by the external environment compared to the other two physiological signals (for example, participants' skin temperatures are quite easy to increase/decrease with the rise/fall of the temperature in their current locations rather than the changes of their emotions). For this consideration, we didn't test the extracted features of Skin Temperature solely but test them with other physiological features instead. From these combinations, we can observe the effects of adding more physiological features on improving the emotion recognition efficiency.

There are various sets of inputs shown on the table below for the multi-classification problems:

Input	
Input-1	BVP + EDA + Skin Temperature
Input-2	EDA + Skin Temperature

Also, there are 3 classes of label outputs:

Label Outputs	
Label-1	Boredom(0)
Label-2	Engagement(1)
Label-3	Anxiety(2)

Classification models:

For all classification models, we used StandardScaler to eliminate the mean (Buitinck et al., 2013)and scaling to the unit variance to further standardise the features.

- **Multilayer perceptron**

  Multilayer perceptron model is a type of feed-forward neural network (Svozil et al., 1997) which is basically consisted of an input layer, hidden layers, and an output layer. We chose it because of their flexibility in learning the mappings from inputs to outputs, as well as its strong self-adaptive and self-learning abilities.

  To solve the multiple classification problem proposed in our research, We trained a MLP model only contains one hidden layer, and there were 200 neurons on the hidden layer. The activation function we used was rectified linear units(RELU), which set the threshold for values at 0 (Agarap, 2018), and used "Adam" for weight optimization because it obtained a better optimization performance than stochastic gradient descent (SGD) in some special case (Zhang, 2018). The maximum number of iterations was set to 800.

- **Random Forest:** As an optimization of Decision Tree, Random Forest is generally seen to have better ability to prevent overfitting in solving such classification problems in our research. Random forests are normally seen as a combination of decision trees, where each tree in the random forest is dependent on the values of an independent random vector sampled with the same distribution (Breiman, 2001). The improvements of random forest comparing to the decision tree were mainly caused by the unit votes of each tree classifier for the most popular input class (Breiman, 2001).

In this task, we set the number of tree in the random forest to 100, and the maximal depth of each tree was 20.

- **Support Vector Machine(SVM):** Due to the advantages of handling classification problems within small sample data and the strong generalization ability, SVM was also trained to classify physiological data into appropriate emotion categories.

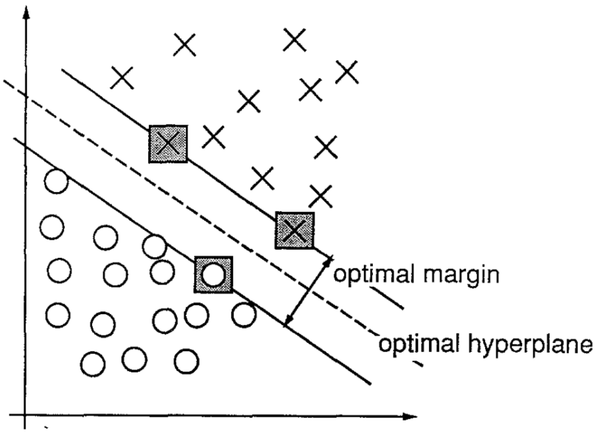

Figure 5.5: SVM models.(Figure from Cortes and Vapnik (1995))

The main implementation thoughts of SVM are to map the input vectors into a high-dimensional features space non-linearly, where a linear decision surface is built with unique characteristics that guarantee the network's excellent generalisation capacity in this space (Cortes and Vapnik, 1995).

- **Logistic Regression** As a common use of the supervised learning approach in machine learning, logistic regression is widely used in solving classification problems like what we proposed in this experiment. To estimate which emotion category a set of extracted features from a physiological signal belongs to, because there are three emotion categories, so we first need to utilise the Softmax function instead of Sigmoid function to map each set of the input data to a number between 0 and 1 and their sum should equal to 1.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K$$

(the formula of Sigmoid function)

Besides, a threshold is set to classify the obtained estimated probabilities.

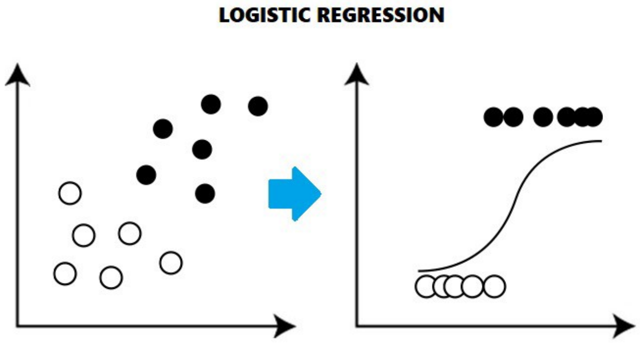

Figure 5.6: example of logistic regression. (Figure from Abhigyan (2020))

The figure above shows the basic mechanism of logistic regression in solving binary classification problems. Following the similar mechanism, to solve the multiple classification problem proposed in this experiment, we directly used the built-in function of sklearn to build the logistic regression model for classification, and further chose One-vs-the-rest (OvR) which consists in fitting one classifier per class as our multi-class strategy sci.

Train and test datasets:

Different from the classification tasks which we implemented before to randomly split training and testing datasets, since this research is designed to gather physiological data and self-rated emotions from 30 participants in 6 game sessions. Therefore, we will train the classification models on the data of the first 24 participants and test how it performs on the data of the remaining 6 participants, which is: we will split training and testing datasets based on participant ID instead of randomly splitting based on a ratio. The aim of this train-test split mechanism was to ensure that emotion recognition could be successfully generalized to new unseen people, instead of only generalizing to different game sessions of the same participants in randomly splitting methods because we need to consider the gaps between multiple users.

Evaluation Metrics:

- **Accuracy:** Accuracy is one of the important and widely-used metrics in evaluating the performances of classification models by calculating the percentage of correct number of predictions among all predictions.
- **F1 Score:** Considering the limitations of accuracy on imbalanced data in the classification task, we adopted F1 score as another evaluation metric. F1 score is

calculated as the harmonic mean of Precision and Recall.

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- **Precision:** demonstrates the percentage of accurate predictions within a given class of predictions.

$$Precision = \frac{\text{True Positive(TP)}}{\text{True Positive(TP)} + \text{False Positive(FN)}}$$

- **Recall:** demonstrates the percentage of examples in a class that the model correctly identified as being in that class.

$$Recall = \frac{\text{True Positive(TP)}}{\text{True Positive(TP)} + \text{False Negative(FN)}}$$

5.3 Statistical Analysis

Several statistical analysis methods were used to solve the research questions proposed. However, before determining which statistical analysis method to use, we first performed normality test to check whether the data is normally distributed, and here we used Shapiro–Wilk test. Moreover, since our sample contained 180 sets of data so the normality can hence be assumed.

Hypothesis testing is one of the traditional and frequently-used approaches in statistical analysis. We chose the ANOVA test to ensure its validity because one of the main goals of our investigation is to determine whether participants experience distinct emotional states when they play Overcooked under varying levels of difficulty. Additionally, each session of Overcooked must be played by each participant twice on the same difficulty setting, in order to determine whether it is easy for them to transition from engagement to a boredom state on the second attempt as they are anticipated to become more accustomed to the game's controls. In this instance, we checked our presumption using a pairwise t-test.

- **Analysis of variance(ANOVA) test:** The ANOVA test was first introduced by an English statistician Fisher, as it separates the systematic and random components of the observed aggregate variability in a data collection, it is utilised by analysts to assess the influence of independent factors on the dependent variable in a regression study. Statistically, the systematic elements but not the random ones have an impact on the data set that is being presented. Kenton (2022).

  Although it is called analysis of variance, ANOVA test is mainly used to compare the variation between the mean values of different groups. Based on that, ANOVA test helps to find out whether the difference of average is statistically significant, as well as revealing whether independent variables affect dependent variables. In our research, we first set the null hypothesis H0: participants play Overcooked in different difficulty conditions won't have different emotional states and an alternative hypothesis H1 as H0's opposite: participants play Overcooked in different difficulty conditions have different emotional states. After that, F value is calculated by the formula below as the test statistic.

$$F = \frac{MST = \text{the mean sum of squares due to treatment}}{MSE = \text{the mean sum of squares due to error}}$$

Then we are able to Calculate the observed value and probability P value of test statistics based on F value, and to make a decision to adopt which hypothesis from the significance level.

- **Paired t-test:** The Student's t-distribution is a family of distributions that are largely identical to the normal distribution, but they appear when the sample data is small and the population's standard deviation is unknown (www). The t-test, also known as a statistical hypothesis test, is based on this distribution. As a type of t-test, paired t-test is widely-used when checking whether there is no difference in the mean of two sets of observations. Furthermore, the paired sample t-test produces pairs of observations since each subject or entity is measured twice.

  In our research, to investigate whether participants were easily to change their emotional states as their gaming skills increase, we need to compare and analysis the data obtained from the first game attempt session and also the second one. Based on that, paired t-test is suitable in answering this research question because the four hard requirements are all satisfied in our experiment:

  1. Observation variables are continuous variables.

  2. Observation variables are independent of one another.

  3. There are no significant outliers in the observed variables.

  4. The difference of observed variables between two matched groups obeys normal (or nearly normal) distribution.

  Similar as ANOVA test, in paired t-test, we also need to set both the null hypothesis and an alternative hypothesis as its opposite based on our research question. However, we need to calculate the test statistic based on the sample mean, sample standard deviation and sample size first:

$$t = \frac{\overline{d} - 0}{\hat{\sigma}/\sqrt{n}}$$

  After that, we can then calculate the probability of observing the test statistic under the null hypothesis to see if there is enough evidence from the results to reject the null hypothesis and accept the alternative one.

Chapter 6

Evaluation

6.1 Survey Analysis

From our experiment design, since we have designed an online survey containing most questions using 5-point Likert scales for the purpose of collecting participants' corresponding feedback after each game-playing session in Overcooked. From data visualization of the collected survey answers from participants, we got some evaluations as below:
Comparisons between different game difficulty conditions

We mainly compared participants' engagement levels, satisfaction levels, cooperation levels with the AI agents and feedback for the two control factors of game difficulties. For each category, we calculated the mean value, standard deviation and median value as various indicators to get more appropriate comparisons.

- **Engagement level**

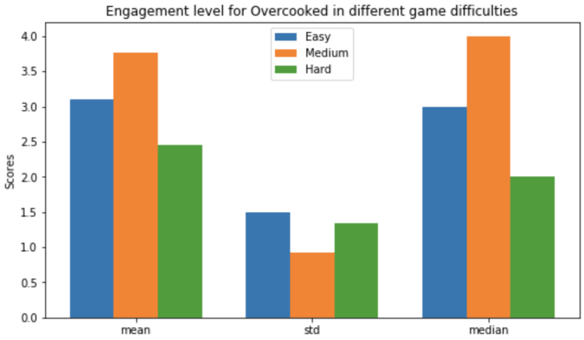

Figure 6.1: engagement levels in different game difficulties.

By looking at the three indicators, it is clear that players of Overcooked in the medium difficulty condition ranked an apparent higher engagement level than in the easy and hard conditions, while players in the hard condition tended to give the lowest engagement scores despite having the highest standard deviation of the engagement level in the easy condition. The overall results are convincing that participants feel most engaged when playing Overcooked in medium conditions.

- **Satisfaction level**

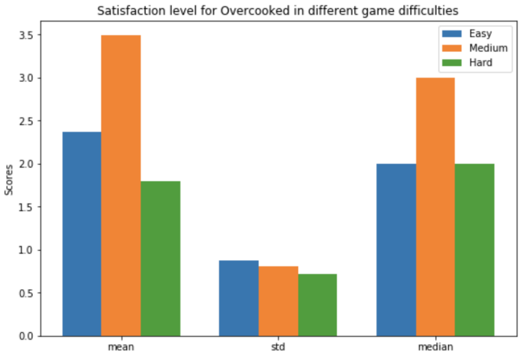

Figure 6.2: satisfaction levels in different game difficulties

According to the average and median satisfaction scores, players who played Overcooked under moderate conditions tended to express the highest levels of satisfaction. Participants playing in the easy conditions had a substantially higher average satisfaction score, despite the fact that the median satisfaction values for the easy and hard situations are the same. The three difficulty conditions' standard deviations for degrees of satisfaction are remarkably similar. Participants reported higher levels of satisfaction with the Overcooked games under medium-difficulty circumstances when all indicators were combined.

- **Cooperation level with AI agent**

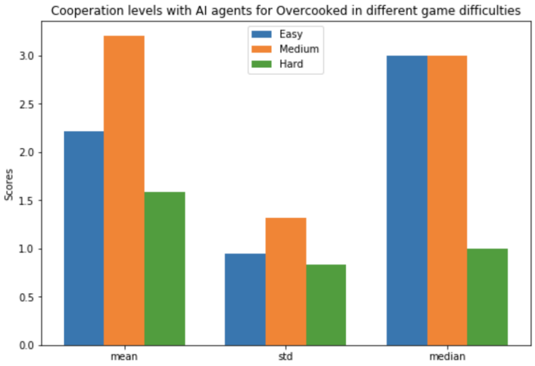

Figure 6.3: Cooperation levels with AI agent in different game difficulties.

Cooperation levels with AI agents in Overcooked are one of the most important factors for us to consider when evaluating user experiences in various game difficulty conditions, we hope that most participants have the highest level of cooperation levels with AI agents in Overcooked when they rate the difficulty level as "Medium", and the corresponding survey answers provide evidence that both the mean and median cooperation levels with AI agents are highest in medium difficulty although there shows the highest standard deviation as well. The results of the easy circumstances come next (even with the same median value as in medium condition but expressing a lower average value), and when players thought the game was quite challenging, they evaluated the AI's collaboration levels as being at their lowest.

- **Game difficulty control factors**

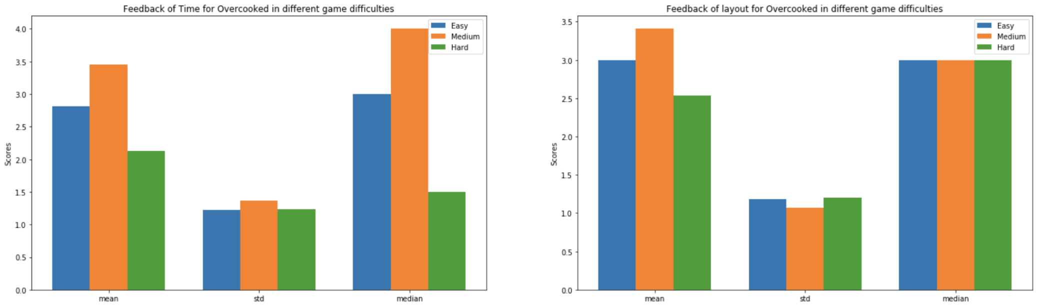

Figure 6.4: Feedback for two game difficulty control factors in different game difficulties.

We want to find out what players' perceptions are of the time duration and game layouts that we used to regulate the difficulty of Overcooked. Actually the standard deviations scores for both factors in all difficulty conditions are quite similar, and the median values for time are the same. We can still observe obvious difference from average values, which indicates that participants tended to give highest scores for the design of time and game layout when they thought the game is in medium difficulty, followed by easy condition, participants gave the lowest scores for both factors when they thought the game is hard.

Comparison between first and second game attempt

In our user experiment, each participants were asked to play Overcooked in same game settings twice, and we hope to observe something between comparisons of both game attempts. From survey answers, we mainly compared their engagement and cooperation levels with AI agents.

- **First game attempt is in Easy condition** When participants rated the difficulty level of their first game attempt as "Easy," we can observe that participants feel less engaged compared to the attempt before from the average engagement levels but expressed unexpectedly higher cooperation levels with AI agents in the

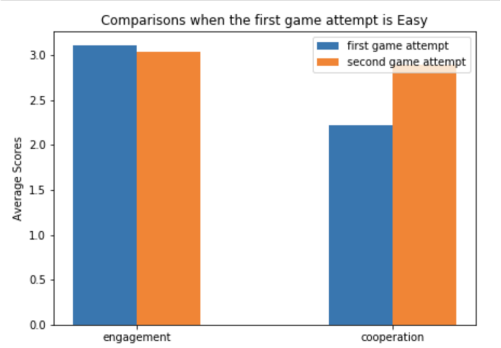

Figure 6.5: Average engagement and cooperation levels for both game attempts(Easy)

second game attempt with the same game settings. We analyzed that it occurred because players felt less engaged (since they had previously believed the game to be easy) due to more familiar game operations, and discovered simpler game methods that could work with AI which results in higher cooperation levels.

- **First game attempt is in Medium condition** When participants rated their first game attempt as "Medium", there are no obvious and huge differences between the average values of both engagement and cooperation levels, but we can still observe that participants felt a bit less engaged and had relatively less cooperation with AI compared to the first one. We deduced that it happened because their user skill increased on the second attempt but the increase is not significant, so they still had similar performances on these two indicators.

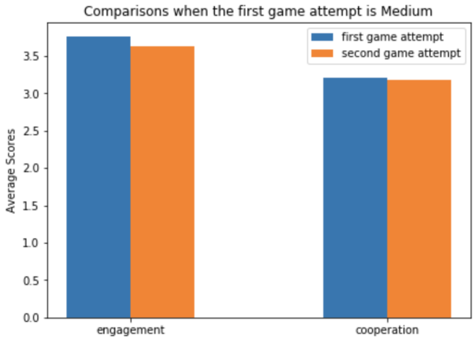

Figure 6.6: Average engagement and cooperation levels for both game attempts(Medium)

- **First game attempt is in Hard condition** When participants rated their first game attempt as "Hard", we can observe obvious differences in both indicators. In the second game attempt, participants presented a significantly higher engagement and cooperation level with AI agents. We inferred that the situation is caused by

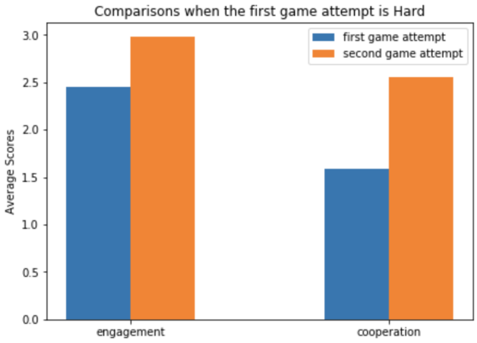

Figure 6.7: Average engagement and cooperation levels for both game attempts(Hard)

the increase in participants' user skills. As their user skill increased, although they found the game challenging on the first attempt, they were able to find optimal game strategies on the second game attempt.

Overall, the evaluation on the collected survey answers almost fits our expectations and assumptions made on it, which is that playing Overcooked in various difficulty condition will results in different levels of engagement, satisfaction, cooperation levels and also cause different ratings on the distinct game difficulty control factors. Also, in the second game attempt, user skills of participants are expected to be increased to the different extent. More specifically, participants are more likely to express highest engagement and satisfaction in medium game difficulty compared to the other two difficulty settings, which makes them having a better cooperation levels with AI agents. Furthermore, as their user skill increase, they felt more engaged and had more cooperation with AI agents when they first thought the game is challenging, while the situation is mostly opposite in the other two game difficulty conditions.

However, due to the limitations of subjectivity of the survey answers, to get a more comprehensive evaluation on how game difficulties vary user game experiences, we also adopted physiological data as an important indicator to explore their specific emotional experiences in varying game difficulties and situations based on growing user skills caused by repeating game attempts.

6.2 Physiological data analysis

Emotion recognition:

As we described in the methodology part, there were three labels indicating different emotional states generated by valence and arousal ratings. However, we eventually found most participants chose the median valence and arousal states in Self-Assessment Manikin(SAM), which made it difficult for us to correctly match their corresponding emotional states in the two-dimensional valence-arousal coordination system. In this case, we added some constraints to define valence equal to 3 as negative and arousal equals to 3 as positive to make emotion matching successful. However, it would increase the imbalanced data because we have added some human interference.

To make the evaluation results more comprehensive, we also tested how different classification models performed in recognizing human emotions by changing the output labels to participants' valence and arousal states. More specifically, it still remained a multiple classification problem, but the labels were switched into two cases:

1. **Valence:**positive valence, natural valence, negative valence.

2. **Arousal:**positive arousal, natural arousal, negative arousal.

Moreover, although we have discussed what features to be extracted from BVP data in the methodology part using one specific BVP data as an example, we eventually found the HeartPy toolkit had set the max iterations automatically, which made it difficult for me to extract the time domain features on whole dataset like PNN20/PNN50 because I failed to change the settings in the original HeartPy. Alternatively, I extracted some statistical features from BVP data with the same data filtering operations based on pyHRV toolkit instead. The new extracted features from BVP were listed in the table below:

Table 6.1: extracted statistical features from BVP

	statistical BVP features
Feature-1	mean value of BVP
Feature-2	standard deviation of BVP
Feature-3	variance of BVP
Feature-4	root mean sqaure of BVP
Feature-5	mean value of the absolute first difference of BVP
Feature-6	standard deviation of the absolute first difference of BVP
Feature-7	mean value of the absolute second difference of BVP
Feature-8	standard deviation of the absolute second difference of BVP
Feature-9	mean value of bpm(beat per minute)
Feature-10	standard deviation of bpm(beat per minute)
Feature-11	mean indices of BVP pulse onsets
Feature-12	standard deviation of the indices of BVP pulse onsets

Comparisons between different labels:

From the results, we can observe that the emotion recognition efficiency was highest when we predicted the valence status (either negative, natural or positive valence), followed by the prediction of arousal status (either negative, natural or positive arousal), and the prediction of emotional states were quite low as we expected. We analyzed it happened

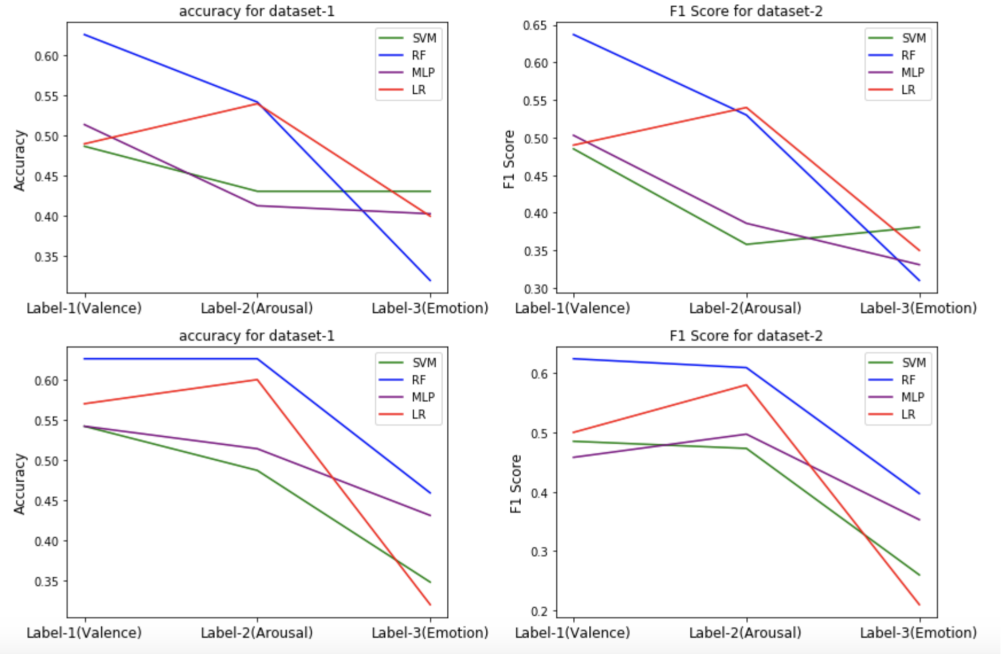

Figure 6.8: Accuracy and F1 Score for different labels(Valence, Arousal and emotion) and different datasets(dataset-1: BVP+EDA+SKT, dataset2: EDA+SKT)

because:

1. Due to the reason that most people chose natural valence/arousal states in Self-Assessment Manikin(SAM), we couldn't match any emotions on each quadrant in the two-dimensional valence-arousal space without adding some constraints to make natural selections to either positive or negative states. In this case, even we successfully matched corresponding emotions, there existed a large bias because the constraints were defined by ourselves instead of participants themselves, the emotions we matched by adding constraints were unlikely to match the correct emotions presented by participants, so it made difficult for us to recognize emotions using physiological data.

2. Compared to using emotions labels, the performances of solving similar multiple classification problems using valence/arousal states had an obvious increase, however, we can still observe that using valence states as labels acquired better performances on most models and datasets. As we discussed before, valence described whether an emotion is negative or positive whereas arousal describes the intensity of emotion. Although we have added detailed explanations for the SAM provided in the questionnaire, most participants still had confusion about choosing the suitable arousal states.

Moreover, from our analysis, for most participants, determining whether their feelings are positive or negative is much easier than determining the intensities. More specifically, especially in playing games, we can easily and quickly judge whether we are happy or not happy, but it might spends more time for us to think about how much happiness or sadness we feel. In this circumstance, due to the differences and difficulties in understanding the concepts of valence and arousal states, when making selections on SAM, the valence states participants chose often reflects their real feeling in playing games, which resulted in a higher accuracy with the matches to extracted physiological features.

Comparisons between different datasets:

We expected that adding more physiological features will make high accuracy in recognizing human emotions, however, we didn't observe an obvious difference between both datasets, where one dataset contained a combination of 24 features selected from EDA, BVP and Skin Temperature data and the other one contained a combination of 12 features extracted from EDA and Skin Temperature data. We analyzed it happened because:

1.Category of extracted physiological features: For BVP and Skin Temperature data, we mainly extracted statistical features like mean value, standard deviation and variance etc. However, statistical physiological features only reflect the pure numerical state but had a low possibility to reflect real human emotions. For example, we cannot directly say that higher EDA data stands for human emotions in higher intensity because the intensity of emotions mostly relies on the number of detected peaks in a series of EDA data, where the feature selection of BVP follows similar rules. However, we mainly extracted peak-related features of EDA but mostly focused on the statistical features of BVP, which means that even we have increased the number of physiological features by adding the extracted features from BVP, but not all of them had strong correlations with human emotions compared to the features extracted from EDA data.

2.Lacking feature selection: Actually, the performances of classifiers in solving such multiple classification problems do not highly rely on the number of features but have higher correlations with the selected features instead. In another word, if we just keep the number of features as many as possible but ignore their correlations with the labels, the performances won't be as good as expected. In our research, to keep a relatively good emotion recognition ability, we have to select those physiological features that have strong correlations with human emotions. More specifically, as the number of peaks and the amplitudes of peaks in EDA and BVP increase, the emotional fluctuations would also increase no matter whether it reflects positive or negative emotion, which means the peaks-related features would have higher correlations with human emotions compared to those statistical features like mean value. We guess the tiny differences between various datasets is because we only considered the number of features but ignored the correlations of features to the labels without performing feature selections.

Comparisons between different classification models:

We have also made comparisons on how different classification models performed on predicting various labels using datasets in same size. From the results, to predict various labels using dataset in the same size, Random Forest obtained the best scores in all three classification tasks, where SVM and MLP obtained a relatively lowest scores. Moreover, when solving the same classification problem using the dataset in different size, Random Forest also performed best for most cases, it obtained a relative better performance not only in the small dataset but also in the large dataset. We analyzed it was because Random Forest is effective in preventing overfitting by introducing randomness.

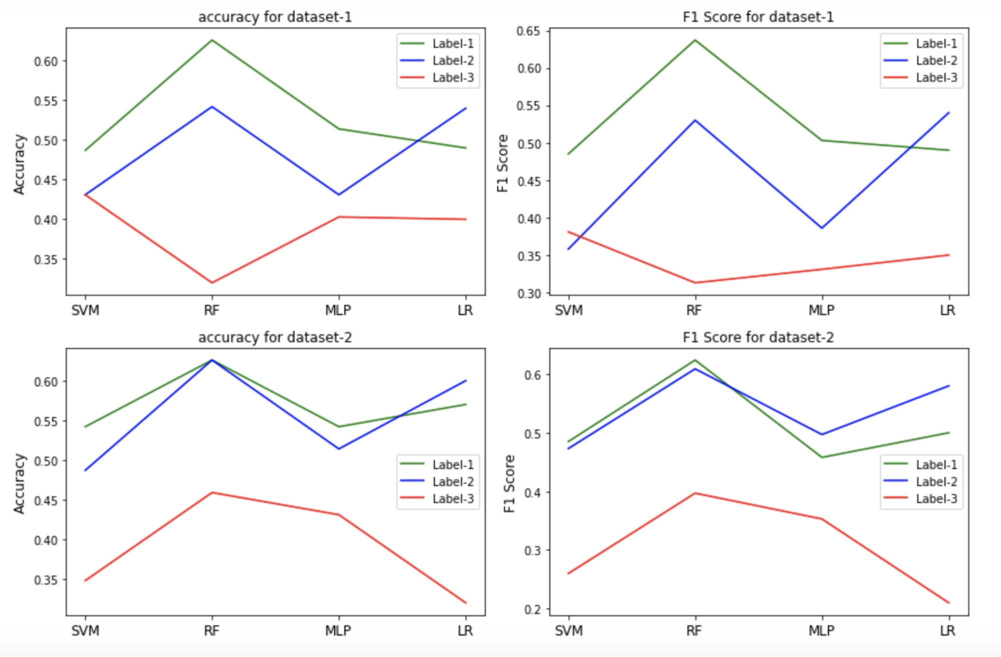

Figure 6.9: Model performances in predicting different labels using different datasets(dataset-1: BVP+EDA+SKT, dataset2: EDA+SKT)

However, we observed that both the highest accuracy score and F1 score were around 0.63 in emotion recognition for all models, and the lowest scores were quite frustrating. We came up with some ideas on the reasons:

1. **Limitations on selected features:**As we have discussed above, we haven't implemented features selection to choose those features with higher correlation as input.

2. **Limitation on the dataset:**Both datasets only contain 180 groups of data but in different dimensions, the data size was a bit small for deep learning models.

3. **Lack cross-validation and model optimizations:** We only tested the performances of these models in simple structure, but we didn't select the optimal parameters for each model, as well as lacking cross-validation.

Table 6.2: the overall results for emotion recognition

	Label-1(Valence state)			
	Dataset-1(BVP+EDA+SKT)		**Dataset-2(EDA+SKT)**	
	Accuracy	F1-score	Accuracy	F1-score
SVM	0.487	0.485	0.542	0.485
RF	**0.626**	**0.637**	**0.626**	**0.624**
MLP	0.514	0.503	0.542	0.458
LR	0.490	0.490	0.57	0.5
	Label-2(Arousal state)			
	Dataset-1(BVP+EDA+SKT)		**Dataset-2(EDA+SKT)**	
	Accuracy	F1-score	Accuracy	F1-score
SVM	0.431	0.358	0.487	0.473
RF	0.542	0.53	**0.626**	**0.609**
MLP	0.431	0.386	0.514	0.497
LR	0.54	0.54	0.6	0.58
	Label-3(Emotion state)			
	Dataset-1(BVP+EDA+SKT)		**Dataset-2(EDA+SKT)**	
	Accuracy	F1-score	Accuracy	F1-score
SVM	0.431	0.381	0.348	0.26
RF	0.32	0.313	0.459	0.397
MLP	0.403	0.331	0.431	0.353
LR	0.4	0.35	0.32	0.21

6.3 Game difficulties and Human emotions

Since we have discussed the usability of physiological data in recognizing human emotions (valence state, arousal state and the matched emotions on two-dimensional valence-arousal space), in this part, we will mainly introduce how various game difficulties relate to human emotions during game playing sessions, as well as answering our research questions and verify our hypothesis.

- **Hypothesis-1:**Participants playing Overcooked in different game difficulties will give rise to different emotions.

  To solve this questions, we performed an ANOVA test on the dataset to check whether participants had emotions in different levels of valence and arousal states in different game difficulty conditions, there was only one variables-game difficulties. From the analysis results, the null hypothesis (H0:changing game difficulties won't make human emotions in different valence/arousal states.) for both indicators (valence and arousal) were rejected: participants felt different valence and arousal states when playing Overcooked in different game difficulties(valence:F=28, p = 1.853420e-11; arousal:F=11.3, P=0.00024). Besides, even the results were convinced enough to show that participants had emotions of different valence and arousal, we still made more analysis on in which game difficulty condition participants felt a a emotion in highest/lowest valence/arousal state:

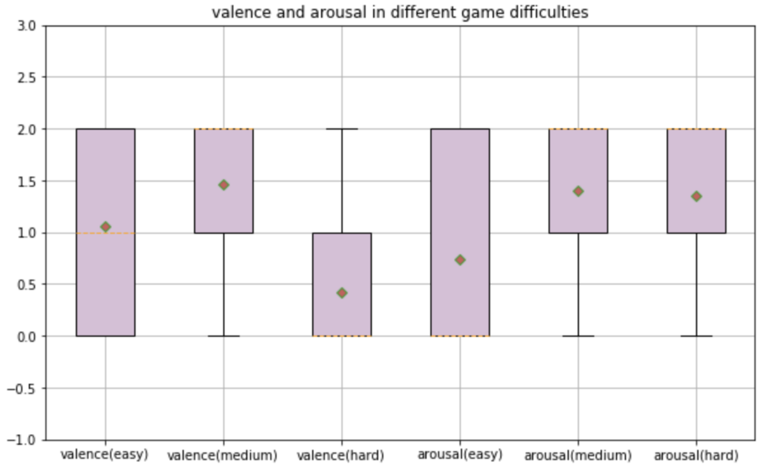

Figure 6.10: mean and standard deviation for valence and arousal states in different difficulties

Valence state:

From the boxplot above, participants played Overcooked in medium condition felt a highest valence, where they felt least valence in hard condition. We analyzed it happened because participants seldom presented positive emotions like enjoyment when they thought the game was challenging, and due to the highest engagement in medium condition they could easily feel happy. Moreover, an easy game might somehow avoid participants from negative emotions.

Arousal state: Participants felt highest arousal in playing Overcooked when they thought the game was in medium condition, but the arousal they felt in hard game was quite similar to the highest one. We analyzed it was because participants were more likely to be aroused as the game difficulties increase due to the reason they thought the game was becoming more challenging and want to get engaged in, where the intensity of their emotion was lowest in easy conditions.

The evaluation results sufficiently proved the importance of dynamic adjusting game difficulties of Overcooked or other cooperative video games in future work, for the reason that participants play Overcooked in easy and hard conditions seldom obtained a better emotional experience, so we have to adjusting game settings to make it neither too easy nor too hard. Moreover, we only examined how user valence and arousal states changed with the changes of game difficulties but haven't check the relationship between the extracted physiological features and game difficulties.

- **Hypothesis-2:**a changes in user emotions will be presented in the second game attempt.

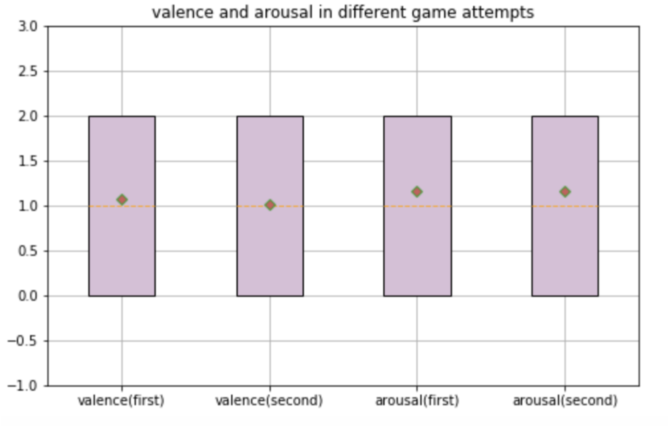

Figure 6.11: mean and standard deviation for valence and arousal states in different game attempts

We performed t-paired test on the datasets to observe whether participants felt an emotion in different levels of valence and arousal in the second game attempt by comparing their valence and arousal states in both game attempts. However, from the results, the null hypothesis(H0: there will be no obvious changes in users' valence/arousal in the second game attempt) was accepted: participants felt similar valence(t=0.43, p=0.67) and also a similar arousal(t=-0.51,p=0.63) in the second game attempt of Overcooked.

From the boxplot reflecting the mean and standard deviation for valence and arousal in different game attempts, we still failed to find any obvious differences for both indicators. We analyzed it happens because of the challenges for participants to increase their game skills in a short time period(in the experiment design, we only set few-minute break between different game attempt), so even they were playing Overcooked in the same game settings for the second time, it was less likely for them to rate the game difficulty into a lower level.

Although our second hypothesis hadn't been proved from the evaluation, we still found some limitations on our experiment design about checking how user emotions will changes with the changes of their game skills. Only setting two repeat game playing sessions had a extremely low probability for participants to improve their game skills.

Chapter 7

Discussion and Limitation

- **Data collection:**

  1. **The number of participants:** We invited 30 participants to play Overcooked in the user experiment. Based on our experiment design, there are nine combinations to test all chosen game layouts and duration, however, to avoid boredom, every 3 participants are invited to test all 9 combinations instead of testing all combinations by one participant. Thus, we have to invite participants as a multiple of 3 as many as possible to get a more comprehensive evaluation, otherwise, there might be bias caused by the limited number of participants.

  2. **Nationalities, ages of the invited participants:** The invited participants are all Chinese students aged around 22. However, the intended users of Overcooked are from all over the world, as well as distributing to all ages and occupations. For this reason, our research is more like exploring human emotional experiences when playing collaborative games for Chinese students rather than more comprehensive intended users of Overcooked.

  3. **Sensor:** The Empatica E4 wristband was used to collect participants' physiological data during game-playing sessions, but the signal detection ability of this sensor is not as precise as those sensors used for medical treatment.

- **Experiment design:**

  1. **Control factors of game difficulties:** We only chose game layouts and duration as two main control factors to distinguish different game difficulties, but the choice of layouts and duration might not sufficient (only 3 choices for each), which might make some participants feel it hard to discriminate various difficulty levels. Also, there might be other factors to control game difficulties in Overcooked.

2. **Participants' preferred choice in questionnaire:** Most questions in our questionnaire use 5-point Likert scales. During the user experiment, although we have already informed participants to try to avoid median choice in advance, most participants still chose median choices for multiple questions, which makes our evaluation difficult and meaningless because we can seldom observe any differences. Moreover, the median selection on SAM increased the challenges to match corresponding emotions in the valence-arousal space, which also resulted in the inaccuracy when evaluating human emotional game experiences.

3. **Time equipped with the sensor:** We set the Empatica E4 wristband to collect physiological data only during the game-playing sessions but not include survey-answering sessions. Which means, although we have set a short break for each participant between each game-playing session to make their physiological signals return to the normal state, their physiological data answering each part of the questionnaire was also useful to reflect their feelings on the corresponding game-playing session, we shouldn't eliminate it.

- **Methodology used in data valuation:**

  1.**Physiological features extraction and selection:** Emotion recognition was one of the important contents in our research, instead of using raw physiological data as inputs to predict human emotions, we needed to extract emotion-related features in advance to ensure the recognition accuracy. However, for the three sorts of physiological data we collected, we only extracted emotion-related features(like peak information reflected the intensity of user feelings) but more concentrated on statistical features for BVP and Skin Temperatures(like mean value, standard deviation etc), those statistical features were not very effective in recognizing human emotions.

  2.**Lack cross validation and model optimization:** In emotion recognition using physiological data, we didn't use k-fold cross-validation. Moreover, we didn't perform any model optimization like adding the size of MLP etc.

  3.**Find relations between human emotions and game difficulties:** After verifying the availability of physiological data in recognizing human emotions during playing Overcooked in changing game settings, we further investigated how human emotions changed in changing game difficulties. However, in the section, we only evaluated the relationship between changing game difficulties/game attempts with valence and arousal state, since they are two major composited indicators to evaluate human emotions, but we didn't explore more about how the physiological features behaves in different conditions.

Chapter 8

Conclusions

8.1 Conclusion

In this research, we explored user emotional experience in playing Overcooked with AI agents under various game difficulties by inviting 30 participants from university students, where difficulties were controlled by the layouts and given time duration, and user emotions were measured by the physiological data collected from the Empatica E4 wristband, as well as their valence and arousal states from Self-Assessment Manikin(SAM). Based on that, we set three game difficulties: Easy, Medium and Hard and three emotional states: Boredom, Engagement and Boredom.

Our evaluation on survey answers proved that participants played Overcooked under Medium condition had the highest levels of engagement and cooperation levels, where their presented relatively lower engagement and cooperation levels both in easy and hard conditions. The similar results were also proved when evaluating their valence and arousal states because participants presented emotions of highest valence and arousal in Medium condition. Besides, we have set repeated game playing sessions to observe whether their emotions and difficulty ratings would be changed due to the increasing game skills. The survey answers illustrated that participants in the second game attempt had obvious different engagement and cooperation levels compared to the first game attempt, but there were not any obvious difference in their valence and arousal states between both game attempts. Moreover, since the self-rated emotions existed strong subjectivity, we also investigated the effectiveness of the collected physiological data in recognizing corresponding human emotions using four different classification models with different combinations of physiological data. The results were convincing to prove its validity, and the emotion recognition ability of Random Forest was best, it received the accuracy of 0.626 and F1 score of 0.637 with the three physiological data(EDA, BVP, SKT) we collected.

The overall results demonstrate the significance of adjusting game difficulties in such cooperative games like Overcooked to maintain high level emotion experiences for intended users, as well as the validity of physiological data in recognizing human emotional experiences.

8.2 Future Work

Our future work will concentrate on dynamically changing game difficulties to make its difficulty levels could always maintain a level that can arouse intended users, as well as inviting more participants from different ages and backgrounds to get a more comprehensive evaluation. Moreover, even though we have proved that users' physiological data could be used to detect their emotions, but the accuracy we received was still a bit low. We will optimize our feature extraction methods to get more emotion-related features from the collected physiological data, as well as optimizing the classification models we used to get a higher prediction accuracy. Also, we still need to improve our mechanism in matching human emotions based on valence and arousal, because the results from our research showed a higher accuracy in predicting user valence and arousal than certain emotions matched in valence-arousal space.

Bibliography

https://www.mu.ac.in/wp-content/uploads/2022/05/Statistical-Methods-and-Testing-of-Hypothesis.pdf. [Cited on page 34.]

Sklearn.multiclass.onevsrestclassifier. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html. [Cited on page 32.]

2021. Ai and data ethics 5 principles to consider. https://www.adp.com/spark/articles/2020/08/ai-and-data-ethics-5-principles-to-consider.aspx. [Cited on page 6.]

Abhigyan, 2020. Understanding logistic regression!!! https://medium.com/analytics-vidhya/understanding-logistic-regression-b3c672deac04. [Cited on page 32.]

Agarap, A. F., 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, (2018). [Cited on page 30.]

Airij, A. G.; Sudirman, R.; Sheikh, U. U.; Khuan, L. Y.; and Zakaria, N. A., 2020. Significance of electrodermal activity response in children with autism spectrum disorder. *Indones. J. Electr. Eng. Comput. Sci*, 19 (2020), 1113–1120. [Cited on page 26.]

Aqajari, S. A. H.; Naeini, E. K.; Mehrabadi, M. A.; Labbaf, S.; Dutt, N.; and Rahmani, A. M., 2021. pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity. *Procedia Computer Science*, 184 (2021), 99–106. [Cited on page 27.]

Bradley, M. M. and Lang, P. J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25, 1 (1994), 49–59. [Cited on page 8.]

Breiman, L., 2001. Random forests. *Machine learning*, 45, 1 (2001), 5–32. [Cited on page 31.]

Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al., 2013. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, (2013). [Cited on page 30.]

Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T.; Seshia, S.; Abbeel, P.; and Dragan, A., 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32 (2019). [Cited on pages 1, 11, 14, 18, 19, and 20.]

Chanel, G.; Rebetez, C.; Bétrancourt, M.; and Pun, T., 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41, 6 (2011), 1052–1063. [Cited on pages 1, 14, and 29.]

Charakorn, R.; Manoonpong, P.; and Dilokthanakul, N., 2020. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. In *International Conference on Neural Information Processing*, 395–402. Springer. [Cited on page 13.]

Chen, S.; Epps, J.; Ruiz, N.; and Chen, F., 2011. Eye activity as a measure of human mental effort in hci. In *Proceedings of the 16th international conference on Intelligent user interfaces*, 315–318. [Cited on page 14.]

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20, 3 (1995), 273–297. [Cited on page 31.]

Du, N.; Zhou, F.; Pulver, E. M.; Tilbury, D. M.; Robert, L. P.; Pradhan, A. K.; and Yang, X. J., 2020. Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation research part C: emerging technologies*, 112 (2020), 78–87. [Cited on page 9.]

Dzieżyc, M.; Gjoreski, M.; Kazienko, P.; Saganowski, S.; and Gams, M., 2020. Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data. *Sensors*, 20, 22 (2020), 6535. [Cited on page 9.]

Fontaine, M. C.; Hsu, Y.-C.; Zhang, Y.; Tjanaka, B.; and Nikolaidis, S., 2021. On the importance of environments in human-robot coordination. *arXiv preprint arXiv:2106.10853*, (2021). [Cited on page 17.]

Frijda, N. H. et al., 1986. *The emotions*. Cambridge University Press. [Cited on page 7.]

Gomes, P.; Margaritoff, P.; and Silva, H., 2019. pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV). In *Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)*, 822–828. [Cited on page 27.]

Herbon, A.; Peter, C.; Markert, L.; Van Der Meer, E.; and Voskamp, J., 2005. Emotion studies in hci-a new approach. In *Proceedings of the 2005 HCI International Conference*, vol. 1. Citeseer. [Cited on pages 2, 13, and 22.]

Jerritta, S.; Murugappan, M.; Nagarajan, R.; and Wan, K., 2011. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th international colloquium on signal processing and its applications*, 410–415. IEEE. [Cited on pages 2 and 8.]

Kenton, W., 2022. Analysis of variance (anova) explanation, formula, and applications. https://www.investopedia.com/terms/a/anova.asp. [Cited on page 33.]

Knott, P.; Carroll, M.; Devlin, S.; Ciosek, K.; Hofmann, K.; Dragan, A. D.; and Shah, R., 2021. Evaluating the robustness of collaborative agents. *arXiv preprint arXiv:2101.05507*, (2021). [Cited on pages 1, 12, and 14.]

Nalepka, P.; Gregory-Dunsmore, J. P.; Simpson, J.; Patil, G.; and Richardson, M. J., 2021. Interaction flexibility in artificial agents teaming with humans. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43. [Cited on pages 1, 11, and 14.]

Stevens, F.; Murphy, D. T.; and Smith, S. L., 2017. Soundscape categorisation and the self-assessment manikin. In *Proceedings of the 20th International Conference on Digital Audio Effects*. [Cited on page 8.]

Svozil, D.; Kvasnicka, V.; and Pospichal, J., 1997. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39, 1 (1997), 43–62. [Cited on page 30.]

van Gent, P.; Farah, H.; Nes, N.; and van Arem, B., 2018. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In *Proceedings of the 6th HUMANIST Conference*, 173–178. [Cited on page 27.]

Van Gent, P.; Farah, H.; Van Nes, N.; and Van Arem, B., 2019. Heartpy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour*, 66 (2019), 368–378. [Cited on page 27.]

Wioleta, S., 2013. Using physiological signals for emotion recognition. In *2013 6th international conference on human system interactions (HSI)*, 556–561. IEEE. [Cited on page 9.]

Zhang, Z., 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, 1–2. Ieee. [Cited on page 30.]

Zhao, B.; Wang, Z.; Yu, Z.; and Guo, B., 2018. Emotionsense: Emotion recognition based on wearable wristband. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 346–355. IEEE. [Cited on page 9.]