Characterizing Physiological and Behavioral Responses Toward Human and AI-generated True and Fake News

Characterizing Physiological and Behavioral Responses Toward Human and AI-generated True and Fake News

Master's Thesis

by

Lian A. Wu

to obtain the degree of

Master of Science in Computer Science

at the Delft University of Technology

to be defended publicly on July 12, 2024
in Delft, the Netherlands

Student Number: 5354285

Thesis Committee:

Dr. Pablo Cesar,	Chair & Supervisor	TU Delft
Dr. Abdallah El Ali,	Daily Supervisor	CWI
Dr. Ujwal Gadiraju,	External Examiner	TU Delft

Contents

SUMMARY

The spread of misinformation on social media has become a prevalent issue, and emerging AI technology further accerlates the generation of misinformation. In this thesis, we investigate how humans perceive AI-generated and human-written news differently and whether they can distinguish between the two. We conducted an experiment that asked participants to evaluate a news dataset consisting of 16 articles with different authenticity (True or Fake) and origin (Human or AI-generated). Physiological signals, including gaze and heart rate data were captured during the study for analysis. The goal was to predict how humans perceive human- and AI-generated news differently based on the collected physiological data. Various data analysis techniques were used to better understand physiological responses and news perceptions. The feasibility of predicting the origin of news, whether it is human- or AI-generated, and whether it is true or fake news based on the user data was assessed. Additionally, we explored how users' general personality and behavioral traits may relate to their ability to classify the news correctly.

1

Introduction

The Internet has become ubiquitous over the last few decades, which fundamentally transformed how information is spread and consumed. While it enables us unprecedented access to information, it opens now challenges due to the facility for spreading misinformation [1]. The way people consume news has shifted towards digital networks, like social media, instead of traditional methods [2]. Misinformation and fake news can have a great impact on society. For example, it was proven that fake news played a role in the 2016 United States presidential election [3]. Studies have also found that there is a clear link between susceptibility to misinformation and vaccine hesitancy, as well as the unwillingness to comply with public health guidelines [4]. The spread of false claims during the COVID-19 pandemic demonstrated how fake news could put public health at risk [5]. The rise in misinformation has become a serious problem. Additionally, the rapid advancement of generative AI has changed how content is created and has paved the way for AI-generated misinformation. Researchers have stressed the importance of understanding how misinformation is received, processed, and shared on social media [4]. For this study, the goal is to gain a better understanding of how humans perceive fake news, including both human-created and AI-generated fake news. Several studies have found a strong link between fake news consumption and eye movement behaviour [6], [7]. It is proposed that the higher volume of saccadic eye movements is due to the increased cognitive load that stems from fake news. However, there is a lack of studies conducted on the emerging AI-generated news.

Previous studies have explored how emotional triggers can lead users to believe in fake news. One finding suggests that emotionally charged content can impair judgment and diminish critical thinking, leading individuals to accept news without further reflection [8], [9]. Emotions are highly complex, and individuals often experience multiple emotions simultaneously. Consequently, accurately classifying emotions typically relies on self-assessment tools, such as the Self-Assessment Manikin (SAM) [10], [11].

Several studies have utilized heart rate variability (HRV) to develop emotion recognition models, albeit with mixed results [12]–[14]. Although HRV has not consistently proven reliable for inferring emotions [14], we have chosen to incorporate heart rate data into our study. This decision is based on the demonstrated correlation between HRV and emotional states, which provides a more objective metric for examining the relationship between emotional responses and the perception of AI-generated misinformation.

1.1. Research Questions and Objectives

The goal of this study is to understand how humans perceive misinformation or AI-generated content. This study will be specifically looking into human behaviour through eye movements and heart rate related responses while reading AI-generated or misinformation content. Furthermore, the correlation between the users' perception of whether news is fake or AI-generated and physiological signals will be investigated. We aim to explore the human perception of AI-generated and falsified content using both the rating responses from the individual and their physiological responses to the news. we try to answer the following research questions:

- RQ1: Do humans perceive human- and AI-generated news differently, and if so, can they correctly categorize the news?
- RQ2: Can eye movements and heart rate data effectively predict users' perception of news' origin (Human or AI-generated) or authenticity (True or Fake)?
- RQ3: Do personality traits, news-seeking behaviour, and trust in technology correlate with the susceptibility of falling for fake news?

To answer the research questions, we will conduct an experiment that gathers participants' news perception and physiological responses to different types of news. Data collected from the eye tracker and heart rate sensor will be processed and analyzed using statistical tests, machine learning modeling techniques.

1.2. Structure of the Thesis

The thesis consisted of six chapters:

Chapter 1: Introduction – Includes the background, significance of the study, and research questions. **Chapter 2:** Related work – Discusses previous research related to eye movements, heart rate variability, fake news, and news datasets. **Chapter 3:** Methodology – Describes the experimental protocol, datasets, questionnaire used in our study, experiment setup, concepts on gaze and heart rate features, and methods of data analysis. **Chapter 4:** Experiments – Discusses the pilot test of the experiment, participant recruitment, experiment procedures, as well as data preprocessing and validation, and the machine learning classifiers modeled on the data. **Chapter 5:** Results – Presents the results of the study, including news perception, statistical tests, and machine learning modeling performance. **Chapter 6:** Conclusion – Summarizes the study, which includes discussion, limitations, and conclusions.

2
Related work

In this chapter, we will introduce the concept of generative AI, explain how it works, and discuss the credibility and linguistic differences of AI-generated text. We will also explore related work on eye movement behaviour, heart rate, and their associations with reading fake news. In addition, we will examine the available news fact-checking sites and news datasets.

2.1. Generative AI

Generative AI, by definition, refers to Artificial Intelligence that can create or produce something [15]. Commonly, the materials generated are in the form of text and images, based on models trained on vast amounts of data. In recent years, the rise of generative AI has been closely linked to advancements in deep learning models, specifically Transformers and GANs (Generative Adversarial Networks) [16]. Transformers [17] excel in text processing, while GANs [18] are powerful in generating images. A well-known example of Generative AI models is GPT (Generative Pre-trained Transformer) by OpenAI.

2.1.1. Transformer

Transformers was first introduced in 2017 in the paper *Attention is All You Need.* [17] Prior to this paper, most state-of-the-art models were built on the Recurrent Neural Network (RNN) architecture. A recurrent model relies on inputting a sentence word by word sequentially. Therefore, a drawback of RNN is that as the sequence gets longer, the memory constrains and suffers from short-term memory. LSTM and GRU can only prolong the memory of RNNs to an extent. Transformers substitute the RNN layer with a self-attention layer. It allows more parallelization in input, and all the sequences can be processed simultaneously. In addition, transformers have no limitation on input length. This is especially useful since context is crucial in language modelling.

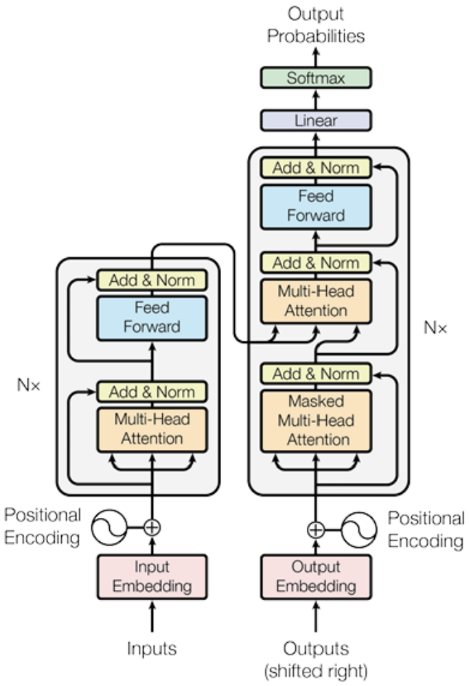


Figure 2.1: Model architecture of Transformer [17]

ENCODER AND DECODER

The encoder and decoder have similar architecture. First, a positional encoding is added to the input and output embeddings. Since the model does not use recurrence, the positional encoding allows transformers to retain the sequence information. Followed by the positional encoding there are stacks of identical layers, with each layer consisting of a multi-head attention sublayer, followed by a feed forward sublayer. The decoder, additionally, has a masked multi-head attention sublayer that takes in known outputs. Residual connection and layer normalization are performed at the end of both sublayers. The encoder encodes the input into a representation with attention so that the decoder knows which word to focus on when decoding [17].

SELF-ATTENTION

The self attention layer allows the model to associate each word to every other word in the input. That is done by mapping *query*, *key* and *value* to an output. The weights on the values are obtained with "Scaled Dot-Product Attention." The formula is as described:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

where d_k is dimension of queries and keys. Compared to the recurrent layer, the self attention layer has the advantages of less computational complexity, parallel computation and the ability to learn long-range dependencies. Multi-head attention enables the model to attend and obtain different representation subspaces. [17]

2.1.2. GPT: GENERATIVE PRE-TRAINED TRANSFORMERS

GPT was developed by OpenAI in 2018 [19]. It uses the Transformer architecture as the foundation of the model. Unlike some models that are tailored for specific tasks, GPT is designed to be able to handle a wide range of language processing tasks.

This versatility is achieved using a traversal-style approach when handling inputs [19]. The structured input sequence is converted into a standardized order that can be processed by the pre-trained model. Their approach reduces the complexity associated with transferred representations of different task-specific models and the need for heavy model customization.

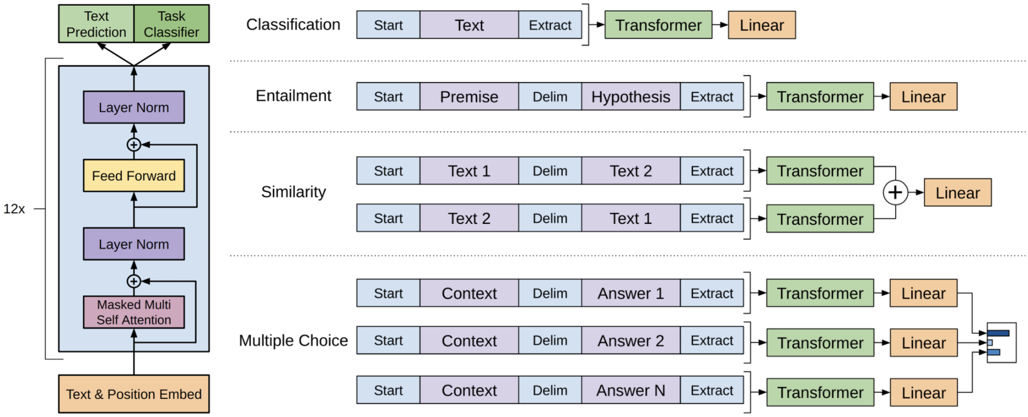


Figure 2.2: Model architecture of GPT [19]. (**left**) training objectives and transformer architecture used. (**right**) The fine-tuning on different tasks of input transformations.

Since the release of GPT-1 in 2018, OpenAI has introduced several new editions. The latest version, GPT-4, was released in early 2023. GPT-3 is trained with approximately 175 billion parameters, making it one of the largest language models at the time of its release. While the exact number of parameters of GPT-4 remains undisclosed, GPT-4 is "more reliable, creative, and able to handle much more nuanced instructions than GPT-3.5," according to OpenAI.[1] In their report [20], OpenAI compared GPT-4 to its predecessor GPT-3.5 and demonstrated a significant performance improvement with GPT-4. It excels in different fields of assessments, can handle more sophisticated tasks while having better accuracy than the predecessors, and also better addresses biases and ethical considerations.

2.1.3. AI-GENERATED NEWS

However, one of ChatGPT's main issues is its ability to accurately generate reliable information. Even though there are significant improvements in GPT-4, OpenAI acknowl-

[1] https://openai.com/research/gpt-4

edges that GPT-4 still retains many of the limitations found in earlier models [20]. Despite the efforts to prevent ChatGPT from being used for illegal or unethical tasks, there remains the concern that GPT-4's enhanced capabilities could pose a greater risk in facilitating such activities than before.

With the rapid advances in generative AI technology, it is becoming more likely that one may encounter AI-generated misinformation. This is concerning because through Generative AI, large volumes of content can be produced in an unprecedentedly short amount of time. In addition, studies have found that humans are not able to determine the genesis of news content [21], [22]; in some cases readers even deemed AI-generated text to be more credible than human-generated text [21]. On the other hand, there are also studies suggesting that AI-generated news headlines are believed less than human creations [23]. Despite this conflicting research, the potential threat of AI-generated misinformation is clear. It is worth noting that the type of Generative AI used in the studies mentioned varied. ([21] with GPT-2, [22], [23] with GPT-3)

2.1.4. Credibility

Graefe et al. [24] conducted a study on human perception of AI-generated news. The study involved more than nine hundred participants who were asked to give ratings on three measures: credibility, readability, and journalistic expertise. The experiment used a 2 by 2 design where users were presented with news labelled as either being written by humans or generated by AI. The study found that when varying the declared origin of the news, subjects showed a preference for news declared as human-written across all three measures, regardless of the actual origin. However, when varying the actual origin of the news, the subjects rated higher credibility and journalistic expertise in AI-generated articles, but found them less readable.

Kreps et al. evaluated the credibility of news generated by GPT-2 [21]. Three versions of GPT-2 were used: small (355M parameters), medium (724M parameters) and large (1.5B parameter). These models produced news articles using 1-2 sentences from baseline *New York Times* articles as input. The AI-generated news used in the study was manually selected after several generation trials with these models. The original *Times* news articles were found to be statistically more credible than the AI-generated news from the smallest model (355M). However, on average, the credibility for the baseline *Times* articles was practically indistinguishable from the news generated by the medium (724M) and large (1.5B) models.

In a different focus on credibility, a study by Yidan Yin et al. demonstrated that AI-generated text responses can create a "feel heard" effect, which validates human feelings [25]. Participants reported feeling "more heard" from AI-generated messages than those generated by humans. However, this effect diminished once the respondents learned that AI was the source. This illustrates that humans tend to trust AI less and appreciate it less when they are aware of its involvement.

2.1.5. Linguistic Difference

Zhou et al. [26] investigated the linguistic difference between AI- and human-generated news. Firstly, they generated misinformation using AI by abstracting core narrative elements from human-created COVID-19 misinformation. With the extracted core narrative prompts, they used GPT-3 to generate the AI version of misinformation that each paired with the human-created one. For their analysis of the characteristic difference in AI-generated misinformation, they considered multiple psycholinguistic features. The results showed that there were significantly more affective and cognitive processing expressions in AI-generated misinformation than human-created ones. Fake news generated from AI was also found to be more likely to cite sources and refer to more testimonial evidence. They discovered that the performance of existing fake news detection models worsens on AI-generated misinformation compared to human creations, likely due to the complexity. However, as generative AI continues to improve, more understanding and research on the topic of AI-generated fake news will be necessary.

2.2. Eye Movement & News Perception

2.2.1. Fixation and Saccades

There are various ways in which fixations and saccades are defined. For example, Larsson et al. characterised that "Fixations are periods when the eye is more or less still, while saccades are fast movements between the fixations that take the eyes from one object of interest to the next" [27, p. 145]. Salvucci et al. define fixations as the pauses over informative regions of interest, and saccades as the rapid movements between fixations [28, p. 71]. Since the eyes are moving so rapidly during a saccade, no new information can be obtained of a saccade under most circumstances [29], [30]. The duration of a fixation and length of a saccade can vary depending on context. Dr. Keith Rayner categorized four tasks: silent reading (reading in silence), oral reading (reading aloud), scene perception (the eyes movements for understanding and interpreting the visual environment [31]) and visual search (the subject is given a target item to find in the visual display and must determine if it is within the display [32]). table 2.1 shows how the fixation duration and saccade length differ according to the tasks.

		SL	
	FD (ms)	Degree	Letters
Silent reading	225-250	2	7-9
Oral reading	275-325	1.5	6-7
Scene perception	260-330	4-5	
Visual search	180-275	3	

Table 2.1: The range of mean fixation durations and the mean saccade length in silent reading, oral reading, scene perception, and visual search [29]. FD = fixation duration; SL = saccade length.

In addition, there are two types of saccades: regressions and progressions. Regressive saccades are saccades that move backwards in the text. It is suggested that when the text is more difficult, there is a tendency to have longer fixations, shorter saccades, and more regressive saccades [29, p. 1460][33]. In his study, the concept of regressive saccades is used to explore how participants may perceive news differently.

2.2.2. Area of Interests

Area of Interests (AOIs) are defined as specific regions within a visual field that are designated for analysis and interpretation of the results in a study. It is essential that these AOIs remain invisible to the participants during an experiment to ensure unbiased responses [34]. In studies involving news stimuli, the AOIs might typically be in rectangle shapes containing the news text and headlines. When comparing the gaze behaviour across different stimuli, researchers typically focus on gaze patterns within the AOIs. This approach helps to minimize noise and concentrate on the most relevant data.

2.2.3. Dwell time

Dwelling time, also known as gaze duration or glance time, is the total amount of time the gaze spends in the AOI—from entering to exiting the AOI [35], [36]. Dwell time is accumulated by each visit to the AOI. This measure can be commonly found in eye movement research.

2.3. Eye Movement & News Perception

In this section, two studies on eye movements and news perception are introduced. Both studies investigated the relationship between fake news and gaze behaviour.

Abdrabou et al. conducted a study on how exposure to fake news affects users' eye movements and mouse movements on the computer [7]. The experiment gathered news content from four categories—Health, Environment, Entertainment and Politics. The posts are text-based, image-based and article-based, which are the formats of posts to be seen on Facebook. The study found that there were significantly more fixations and longer time spent on a post when the participants were reading fake news. The average fixation counts and duration have a big impact on the classification accuracy. They suggested that the behaviour difference while reading fake news is due to the induced cognitive load when processing fake news. The six main extracted eye movement features were fixation count, average fixation duration and distance, average saccadic duration and length, and gaze duration in one post. The fixations were classified using the Dispersion-Threshold Identification algorithm [28]. The three machine learning techniques—SVM, Logistic Regression, and Random Forest—were used for the classifiers. SVM gave the best result in most cases. The classifier using gaze features only achieved the highest prediction rate of a 64.2%. It performs better than the models using mouse features only and the combination of gaze and mouse features.

Sumer et al. [6] conducted the study FakeNewsPerception. The study showed that participants spent significantly more time reading fake news than real news, and exhibited more fixations and saccades on the fake content. A further analysis study [37] was performed on the FakeNewsPerception dataset. It further classified saccades as regressive or progressive. They found that while reading fake news, there were significantly more regressive saccadic eye movements observed than when reading the real news. They also suggested that even when the truthfulness of the news was not known to the people, the visual behaviour differed depending on whether the content was fake news or not.

The eye movements data in the FakeNewsPerception experiment was captured with a Tobii Spectrum eye tracker. In total, 27 people participated in the study. After going through eye tracking calibration, participants were instructed to read 60 news items. The news was ordered randomly. Each news item had a true and false versions. The authors ensured the two versions appeared the same number of times by reversing the version for every other participant. In the second part, the participants got 10 seconds to rate the credibility of the news on a 5-item Likert scale. Afterwards, the CRT (Cognitive Reflection Test) of the participant was measured, followed by a questionnaire on News-Finds-Me perception and political orientation. The data was pre-processed with the Tobii Pro-Studio. The pupilometry data was also collected during the experiment.

The dataset collected from the experiment was then made available for future research to better understand how humans perceive fake news, with a focus on the link between eye movements and the perceived news truthfulness. It contains data on eye movements while reading, the perceived believability scores, political orientation, cognitive reflection test, and News-Find-Me perception [6]. The processed data includes aggregated features such as fixation information, saccades per participant, and stimulus. The stimulus data contains annotated regions of interests, headlines, source of media, etc. It is suggested that the data can be used to classify the perceived believability, or, in contrast, to classify whether the news is true or false.

2.4. Personality, Psychology, and News Credibility

Various traits have been reported to be associated with someone having a higher tendency to believe fake news. Studies have shown that a lack of critical thinking [8] or a reliance on emotion [9] may result in someone falling for fake news. Personality traits can play a role, too. A study in 2018 [38] found that based on the predicted Big 5 personality types, people that are more extroverted and agreeable are more likely to trust fake news.

2.5. Heart Rate Variability & News Perception

Heart rate variability has been used to reflect on cognitive abilities, personality traits and neural processes [39]. Heart rate, also known as beat per minute (bpm), can be an indicator of stress response when heart rate increases. Heart rate variability (HRV) on the other hand is the variation in time intervals between heartbeats. The short-term changes of HR are affected by the regulatory mechanisms of the baroreceptors, which are part of the autonomic nervous system (ANS) [39], [40]. Therefore, when blood pressure increases, the barorceptors trigger to increase the diameter of blood vessels to decrease the HR. This phenomenon is called a baroreflex and is linked to heart rate variability. A higher HRV indicates a responsive ANS, which shows that the body can manage and react to stress and relaxation efficiently.

The autonomic nervous system can be further divided into two units: the parasympathetic nervous system (PNS) and the sympathetic nervous system (SNS). The PNS is also known as the "relaxed response" system that predominates at rest, while SNS is known as the "quick response" system and predominates during stressful states [39].

Kirkwood et al. used HRV and skin conductance for the study on the perceived believability of news headlines from social media [41]. An appetitive system response is derived from the heart rate and skin conductance data. The study investigated whether participants showed an appetitive or aversive response based on whether they believed that the news was true or false. They found that fake news was correlated with the appetitive response more than with true news, which resulted in higher HRV and skin conductance. The experiment was carried out with 26 participants. They gave ratings on the believability of 50 sets of news headlines paired with an image. Each news set was shown for 15 seconds.

There have been studies on the correlation between reading clickbait news headlines and emotional arousal [42]. It was shown that a higher level of emotional arousal was observed in clickbait news headlines compared to neutral phrasing headlines. Clickbait headlines are generally written in a way that is misleading and sensationalized. It shows that news headlines could be fabricated and induce emotional arousal from the readers. Another finding from the study is that SAM and pupillary dilation is relatively consistent.

2.6. FAKE NEWS DATASETS

Various fact-checking websites take efforts to identify misinformation and verify its legitimacy from news sources. PolitiFact[2] is one that is widely used for verifying the claims of public figures. It is a platform that offers in-depth analysis reports on the news and references links for each case. Snopes[3] is another fact-checking site that is known for debunking myths and viral news stories. Many fake news datasets were compiled from PolitiFact. A well-known one being LIAR [43], which contains more than 12.8 short statements that were labelled manually in different context retrieved from PolitiFact. The dataset FakeNewsNet [44] also includes news from GossipCops, which fact-checks rumors and claims made about celebrities. There are other online platforms that focus on different topics of interest in news [7]. For example, ClimateFeedback[4] and HealthFeedback[5] provide news analysis in the areas of climate and health [7].

However, there are not many publicly available AI-generated fake news datasets. The AI-generated misinformation dataset is made available from the study Synthetic Lies [26]. Therefore, her study design and selection of other stimuli are based on this dataset. The scarcity of AI-generated news datasets could be attributed to the recent emergence of Generative AI technology, which is still new and rapidly evolving. Additionally, the generation of misinformation with AI is unethical, leading to more precautions in data sharing [26].

[2] www.politifact.com
[3] www.snopes.com
[4] https://climatefeedback.org/
[5] https://healthfeedback.org/

3
Methodology

3.1. Protocol

The experiment is divided into two parts: News Rating and Survey. The news stimuli used for the study were derived from the results of the pilot test, which is described in the next chapter in section 4.1.

Participants will be able to read through a selection of news stories at their own speed, while their heart rate and eye movements are monitored. After reading each news article, they will be asked to rate its believability. The news stories presented will be a combination of real and fake news. Finally, participants will be asked to complete a questionnaire to determine their personality type [45], their familiarity with the news they read, their cognitive reflection test, their Propensity to Trust in Technology, and their News-Finds-Me perception.

3.2. Datasets and News Categories

The available dataset is discussed in section 2.6. The datasets used for the stimuli are COVID19-FNIR [46] and AI-Misinfo [26]. Of the 40 stimuli, 20 come directly from COVID19-FNIR, 10 were rewritten by GPT-4 from CONVID19-FNIR, and the other 10 come from the Synthetic Lies dataset [26]. There is an equal number of true/fake and human/AI-generated news in the dataset. All 40 news will be used in our pilot test. To ensure uniformity of the dataset, only the news headline and body text are displayed to the participants in the experiment. The font formatting is kept uniform, and any media published with the article is removed. The publication date is also omitted, as date information is not provided in AI-misinfo. In several occasions, the content of the news article contained the publisher's name. We chose to present the news as close to its origin state as possible and did not omit publisher information when it was self-disclosed in the article. Additionally, studies have shown that publisher information poses no significant effect on whether participants perceived the news as true or fake [47]. The four categories of news are as follows:

- Real-Human: Real news written by humans from the COVID19-FNIR Dataset

- Fake-Human: Fake news written by humans from the COVID19-FNIR Dataset
- Real-AI: Real news generated by GPT-4, true news from COVID19-FNIR as inputs
- Fake-AI: Fake news generated by GPT-3 from the Synthetic Lies dataset. The 10 news articles with the highest word counts were used for the studies –around 200 to 300 words each.

3.2.1. Human-created News: COVID19-FNIR

The Human-created stimuli came directly from the COVID19-FNIR (section 3.2.2). The dataset contains both True and Fake news. Each True News is provided with the original web link to the article. Each Fake News comes with a link to a fact-checking website, which explains why the news is fake in detail. The original URL and a web archive link for the fake news article can also be found. In many cases for Fake News, the original URL to the fake news no longer works. However, the news text can still be extracted using the web archive link. The main criterion for selecting stimuli is the word count of the news articles.

3.2.2. AI-generated News

True: COVID19-FNIR and Synthetic Data Creation

The AI-generated True News stimuli in the study were generated with the state-of-the-art language model GPT-4, developed by OpenAI. To produce AI-generated true news, we initially selected human-written true news from the COVID19-FNIR dataset. The main criteria of the selection are 1) the length of the articles and 2) the topics of the news. Our goal was to cover a wide variety of subjects while ensuring that the topics are accessible to a broad audience with diverse backgrounds. We used the OpenAI API to generate news articles, ensuring independence from previous chat histories by using a unique token for the API. The configuration parameters used are `engine='gpt-4-0613`, and `temperature=1`. The `temperature` can range from 0 to 2, which controls the model's randomness [1]. We opted for the default value 1 to strike a balance between the trustworthiness of the text and the flexibility in the writing style of the generative AI model. Subsequently, we manually reviewed the generated texts to verify their authenticity based on the original articles, while ensuring that the length of the text is approximately between 200 and 300 words. The process was repeated for one true human news input until the AI-generated news text met both criteria.

For reference, the following prompt was used to generate the news text:

> **Prompt to Generate News Text**
>
> "Please rewrite the provided news article, relying solely on the information given from the news article, and write it in the style of a news article. The text should be strictly between 200 - 300 words, and please also generate a news headline based on the news article."

[1] https://platform.openai.com/docs/api-reference

Fake: Synthetic Lies

Zhou et al. [26] used GPT-3 to produce AI-generated misinformation, the so-called "AI-misinfo" dataset, which contains 500 pieces of AI-generated misinformation. There are two types of misinformation in AI-misinfo dataset: News and Non-news. News is when formal reporting of events or matters is done; Non-news is when informal communication is used to report on information or topics, similar to a post on a social media platform. For this study, only the "News" type of misinformation is used. The average length of the "News" in AI-misinfo is 116.01 tokens. To standardize the length of the stimuli for the study, the top ten news stories with the highest word count were chosen. The average word count of the ten stories was 235 (with a standard deviation of 32, a maximum of 307, and a minimum of 199). The news headlines were not included in the AI-misinfo dataset. Hence, the news headlines were generated using GPT-4 with version `gpt-4-0613`, with the news text as input. Using the prompt:

> Prompt to Generate News Headlines
>
> "Please generate an appropriate news headline for the news article provided."

3.3. Heart Rate

3.3.1. Heart Rate Variability

For the study we look at Heart rate, and some difference-based indices, including SDSD (standard deviation of the successive difference), RMSSD (root mean square of successive difference). Based on the previous studies, RMSSD is more preferred because of statistical robustness [39], [48]. Heart rate variability is measured with an Empatica 4 device in this study. For the analysis of the data we use the python library called Neurokit. The blood volume pulse (bvp) is captured at a rate of 64 per second. For the study, we use both HRV_RMSSD and PPG_HR processed from Neurokit.

$$\mathrm{RMSSD} = \sqrt{\frac{\sum_{i=1}^{N-1}(RR_i - RR_{i+1})^2}{N-1}} \tag{3.1}$$

3.3.2. Electrodermal Activity

Electrodermal Activity (EDA) is the measure of skin conductance. It changes according to moisture level of the skin, which is influenced by the sweat gland activity [49]. The higher levels of electrolytes in sweat lead to a greater skin conductance [50]. The sympathetic nervous system is closely related to the variation of the EDA signal level [49]. Therefore, in physiological research, EDA is used to assess the ANS activity, especially in the sympathetic nervous system which controls the fight or flight responses. EDA can be split into two components:

- Tonic EDA: Tonic EDA is the baseline level of skin conductance. It varies slowly overtime. It is measured as the skin conductance level (SCL).

- Phasic EDA:Phasic EDA provides the rapid changes in skin conductance. The changes in phasic EDA often occur within seconds to minutes. It is also known as skin conductance responses (SCR).

3.4. Eye Tracker Calibration

The two key things considered in eye tracking are precision and accuracy. Accuracy is described as the average difference between the actual position of the stimuli and the captured gaze position. Precision, on the other hand, refers to the eye tracker's capability of precisely replicating the same gaze point measurements. Precision assesses the variation of gaze point data collected through the Root Mean Square (RMS) of successive samples [35]. From equation (3.2), θ represents the degree of the visual angle. In addition, Precision can be evaluated using SDPrecision, which is the standard deviation of the normalized RMS. Figure 3.1 illustrates the various scenarios in combination with good or poor precision and accuracy.

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\theta_i^2} = \sqrt{\frac{\theta_1^2 + \theta_2^2 + ... + \theta_n^2}{n}} \tag{3.2}$$

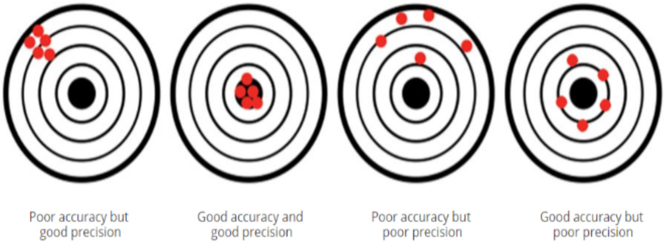

Poor accuracy but good precision
Good accuracy and good precision
Poor accuracy but poor precision
Good accuracy but poor precision

Figure 3.1: The figure demonstrates the results of a combination of poor and good accuracy or precision

3.5. Setup

In this section, the components of the experiment, the functionalities of the hardware, and the software implementation details are introduced. The main components of the experiment systems are the eye tracker, Tobii Pro Fusion, the heart rate sensor, Empatica E4, the study's proprietary web app, PhysioNews, and the software Tobii Pro Lab.

3.5.1. Hardware: Tobii Pro Fusion

Tobii Pro Fusion[2] is used to measure eye movements in the experiment. It is a screen-based eye tracker that is tolerant to head movement. It has a dual-camera system which captures eye data more accurately than with one camera. The eye tracker captures eye movements up to a rate of 120 images per second. It is placed under a monitor. The eye data recorded includes the gaze, eye openness, and pupil information.

[2] https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion

Figure 3.2: The ease of set up of the Tobii Pro Fusion [51] – it can be easily connected to the computer and attached to the monitor when in use.

3.5.2. HARDWARE: EMPATICA E4

Heart rate data will be collected using an Empatica E4 wristband[3]. E4 is a wireless wearable device continuously captures the data from the wearer. The main functionalities are using photoplethysmography to collect blood pulse, heart rate, heart rate variability data as well as measuring the electrodermal activity. The PPG sensor has a sampling frequency of 64 Hz.

The E4 device can be connected via Bluetooth to the E4 Realtime mobile app. Using the app we could view the participant's visualized data in real-time and check the status of the device.

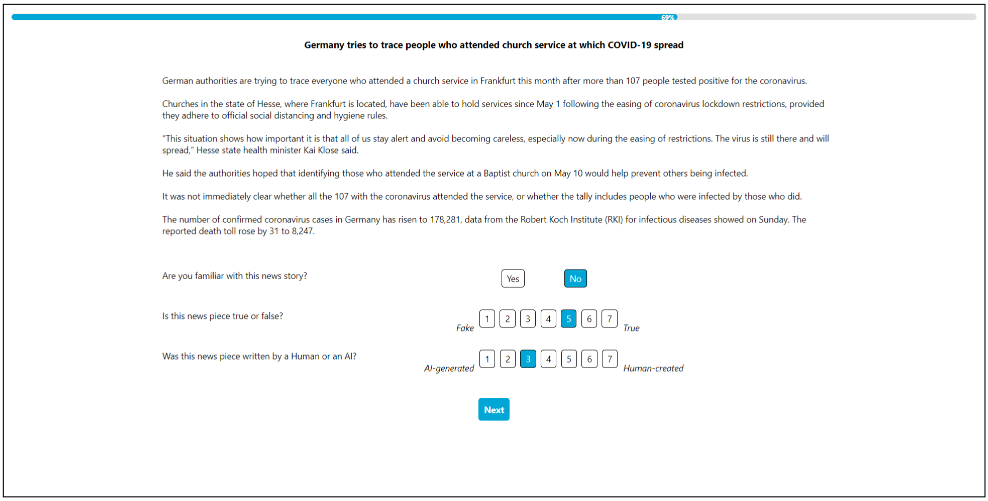
Figure 3.3: The Empatica E4 [52] used for the study.

3.5.3. WEB APP: PHYSIONEWS

PhysioNews is a web app that was created exclusively for this experiment. The app consists of five main parts: an introductory page, a page for collecting demographic information, a section for news stimuli, a section for survey questions, and a page that displays the results of correctly classified news. For the front end, we used React.js. For the backend, Node.js with Express and a REST API were used. The site and database are hosted on Heroku, with ClearDB serving as the cloud server for the MySQL database.

[3] https://www.empatica.com/research/e4/

During the experiment, GET requests are used to retrieve news stimuli and survey questions data, while POST requests are used to record participants' ratings of the news stimuli and their survey responses in the database. A progress bar is implemented to help participants track their progress, as shown in figure 3.5a, indicating how much work has been completed and how much remains.


69%

Germany tries to trace people who attended church service at which COVID-19 spread

German authorities are trying to trace everyone who attended a church service in Frankfurt this month after more than 107 people tested positive for the coronavirus.

Churches in the state of Hesse, where Frankfurt is located, have been able to hold services since May 1 following the easing of coronavirus lockdown restrictions, provided they adhere to official social distancing and hygiene rules.

"This situation shows how important it is that all of us stay alert and avoid becoming careless, especially now during the easing of restrictions. The virus is still there and will spread," Hesse state health minister Kai Klose said.

He said the authorities hoped that identifying those who attended the service at a Baptist church on May 10 would help prevent others being infected.

It was not immediately clear whether all the 107 with the coronavirus attended the service, or whether the tally includes people who were infected by those who did.

The number of confirmed coronavirus cases in Germany has risen to 178,281, data from the Robert Koch Institute (RKI) for infectious diseases showed on Sunday. The reported death toll rose by 31 to 8,247.

Are you familiar with this news story? Yes [✓ No]

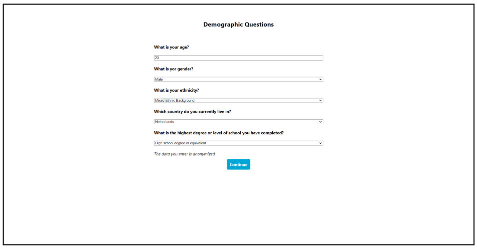
Is this news piece true or false? *Fake* 1 2 3 4 [✓ 5] 6 7 *True*

Was this news piece written by a Human or an AI? *AI-generated* 1 2 [✓ 3] 4 5 6 7 *Human-created*

Figure 3.4: Web app: PhysioNews — News Rating section

3.5.4. Software: Tobii Pro Lab

The Tobii Pro Lab software allows users to set up experiments and export the results. The Tobii Pro Fusion eye tracker is used in combination with Tobii Pro Lab. The software allows researchers to use forms of materials as stimuli, including text, images, videos, or web pages. In this case, the PhysioNews web app was used as the stimulus. This novel implementation of the web app and survey grants the ability for rich customization of the experiment system. For instance, easily record survey responses to the database. Tobii Pro Lab also plays a key role in processing and interpreting the raw eye movements


Demographic Questions

What is your age?
23

What is yor gender?
Male

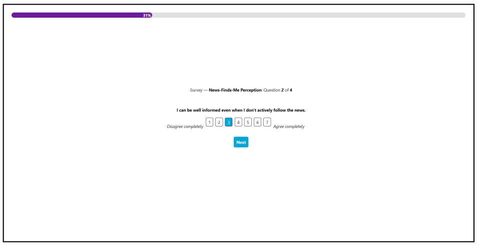
What is your ethnicity?
Mixed Ethnic Background

Which country do you currently live in?
Netherlands

What is the highest degree or level of school you have completed?
High school degree or equivalent

The data you enter is anonymized.

(a) Web app: PhysioNews — demographics screen


31%

Survey — **News-Finds-Me Perception**: Question **2** of **4**

I can be well informed even when I don't actively follow the news.

Disagree completely 1 2 [✓ 3] 4 5 6 7 *Agree completely*

(b) Web app: PhysioNews — Survey section

Figure 3.5: Screenshots of the PhysioNews web app

data collected through the Tobii Pro Fusion eye tracker. The raw gaze points are grouped into Fixation or Saccade with the Tobii I-VT (Fixation) filter. I-VT is a velocity based algorithm for identifying fixations and saccades [28].

3.6. Feature Extraction and Data Processing approaches

3.6.1. Gaze Features

Gaze within Areas of Interest (AOIs) is considered. Definitions are as follows:

- Gaze duration in AOIs: The total time of gaze duration within AOIs.
- Saccade count in AOIs: The number of whole saccades occurring within AOIs.
- Average saccade length: The mean length of saccades measured within the AOIs.
- Average fixation duration: The average duration of a fixation, including both partial and whole fixations.
- Average pupil diameter of Fixations: The mean diameter of the pupil during fixations.
- Pupil diameter: The mean diameter of the pupil throughout the Time of Interest (TOI).

3.6.2. HR+EDA Features

Baseline data are derived from values during the introduction and demographic screens. Definitions are as follows:

- Beat Per Minute (BPM) mean varying ratio to baseline: The ratio of change in time between BPM compared to baseline.
- RMSSD varying ratio to baseline: The ratio of change in the root mean square of successive difference (RMSSD) compared to the baseline value.
- EDA Tonic varying ratio to baseline: The ratio of change in the time between EDA Tonic compared to the baseline.
- EDA Phasic varying ratio to baseline: The ratio of change in EDA Phasic compared to the baseline value.

3.6.3. Tobii I-VT Filter

The Tobii I-VT filter groups eye movements data into Fixation or Saccade.

3.6.4. Savitzky-Golay Filter

The Savitzky-Golay filter[53], [54] is a data smoothing technique used in signal processing to reduce noise in a signal. This can be applied to the data of heart rate and diameters of pupils in our study. It is a type of low-pass filter that uses a polynomial fit to the data

Algorithm 1 Pseudocode for the I-VT algorithm.

```
procedure I-VT(protocol, velocity threshold)
    Calculate point-to-point velocities for each point in the protocol
    Label each point below velocity threshold as a fixation point,
        otherwise as a saccade point
    Collapse consecutive fixation points into fixation groups removing saccade points
    Map each fixation group to a fixation at the centroid of its points
    Return fixations
end procedure
```

points within a window to smooth out the signal. equation (3.3) shows the Savitzky-Golay smoothing while polynomial order $N = 0$ and $M = 1$.

$$y[n] = \frac{1}{2M+1} \sum_{m=n-M}^{n+M} x[m] \tag{3.3}$$

3.7. Approach to Statistical Testing and Analysis

Mann-Whitney U and Wilconxcon Signed-Rank Test are used. Both are non-parametric tests, that they do not assume the data follows a normal distribution. Therefore, the tests are less sensitive to outliers than parametric tests. The main difference between the two tests is that Mann-Whitney U is paired, while Wilcoxon Signed-Rank test is on two independent groups. For our study, it is a paired setting where participants go through the same set of stimuli of different categories [55].

3.8. Survey

In order to gain understanding of the participants in the experiment. Several questions were included at the end of the experiment. The questionnaire also helps test whether an individual's personality, news seeking behaviour, or their trust in technology contributes to their likelihood of categorizing a stimulus as true/fake and human/AI-generated. The survey questions can be found in the Appendix.

General Questions

The general questions assess the participants' level of English and how closely they follow American political news and COVID-19 news. The information provides more context of the participants' backgrounds.

News-Finds-Me Perception

The News-Finds-Me perception test[56] investigates if the individuals believe that they can stay informed without actively seeking out news. The survey is not directly link to the motivation behind seeking news, but if focuses on whether individuals hold the belief that news will find them. The survey is also used in the study from Sumer et al. [6].

PROPENSITY TO TRUST IN TECHNOLOGY

Propensity to Trust in Technology (PTT) is a test to evaluate an individual's attitude towards technology and their trust in it [57]. The measure, developed by Schneider et al., consists of 6-items [58]. Participants are asked to rate their agreement with each prompt on a scale of 1 (strongly disagree) to 5 (strongly agree). This experiment wishes to explore if one's trust in technology, in this case in AI-generated news articles, would affect their trust in the authenticity of the news article.

COGNITIVE REFLECTION TEST

Cognitive Reflection Test (CRT) is a measure consists of three items that was first proposed by Frederick in 2005 [59]. The tasks can be easily understood, however the suppression of a initial wrong, intuitive answer is needed to derive the correct answers. The aim is to assess this inclination to resist the initial incorrect response, and to engage in further reflection that ultimately result in the correct answer. [59], [60]. The CRT items can be found in section 6.2 The survey section of the experiment, started off with the cognitive reflection test (CRT). The test immediately follows the last stimulus from the news rating section. This helps to see the cognitive state of the participants right after going through the stimulus.

THE BIG FIVE PERSONALITY TEST

The Big Five Inventory (BFI) is a widely used personality test that evaluates one's personality [61]. For our study, we use the BFI on a 5-likert scale. The five key traits of BFI are Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness. When BFI was first introduced in the late 1980s, it was with a total of 44 items. Studies later have shown that a shorten version of 10-item BFI retains its validity and reliability [62]. The compact format is particularly beneficial where time efficiency is important. Given that the experiment already includes multiple surveys, the 10-item version BFI was chosen.

4

Experiments

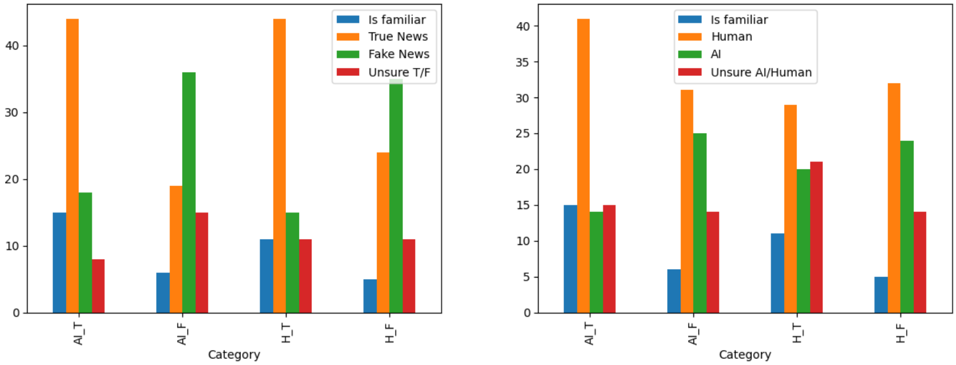
4.1. Pilot Test

A pilot test was conducted to gain an understanding of how people perceive the news stimulus used in the study. The participants in this test were asked to review 40 news articles. For each of the news articles, they were asked three questions. First, to rate if they think the news is true news or fake news. Second, to rate if they think the news is AI-generated or human-created. For these two questions, participants also have the option to choose unsure if they do not know how they would classify the news. Lastly, a third question was included to ask if they are familiar with the news stimulus. This helps determine whether their judgement is affected by their previous knowledge of the news piece.

4.1.1. Results

The table below shows the count of labelled of whether the news is true/fake, human/AI. The first table breaks down the data from the 2 × 2 categories (N=70). The second table shows the counts based on one feature (N=140).

In total, seven people participated in the pilot study. The results can be seen in figure 4.1. In terms of whether the news is true or fake, participants could identify them more easily. Regardless whether the news is human or AI-generated, we observed similar percentage of labelling the news as true and fake. 62.8% of True news were labelled as True, while 50.7% of fake news were labelled as Fake. There are more fake news mislabelled as true news than vice versa, but also more unsure responses were given when it is a fake news (18.6% unsure in fake news, 13.6% unsure responses in true news).



(a) True or Fake

(b) Human or AI

Figure 4.1: Pilot Test Result N=7

	labeld as					
Type	True	Fake	Unsure	Human	AI	Unsure
AI-T	**62.8% (44)**	25.7% (18)	11.4% (8)	**58.6%(41)**	20% (14)	21.4%(15)
AI-F	27.1% (19)	**51.4% (36)**	21.4%(15)	**44.3% (31)**	35.7% (25)	20% (14)
H-T	**62.8% (44)**	21.4% (15)	15.7%(11)	**41.4% (29)**	28.6% (20)	30% (21)
H-F	34.3% (24)	**50% (35)**	15.7% (11)	**45.7% (32)**	34.3% (24)	20% (14)
True	62.8%(88)	23.6%(33)	13.6%(19)	50%(70)	24.3%(34)	25.7%(36)
Fake	30.7%(43)	50.7%(71)	18.6%(26)	45%(63)	35%(49)	20%(28)
Human	48.6%(68)	35.8%(50)	15.7%(22)	43.6%(61)	31.4%(44)	25%(35)
AI	45%(63)	38.6%(54)	16.4%(23)	51.4%(72)	27.9%(39)	20.7%(29)

Table 4.1: Distribution of labelled news: Pilot test

4.1.2. FEEDBACK

The primary feedback received from the participants indicated that reviewing 40 news articles was overwhelming. A common suggestion was to reduce the number of articles. Many reported difficulty maintaining their concentration throughout the entire questionnaire. Some participants expressed frustration that a few of the news articles were time-sensitive, making it challenging to judge whether they were true or fake without knowing the time of publication. For instance, a news story about a celebrity contracting COVID-19 might be false at the onset of the outbreak, but could later become true. Additionally, participants expressed the wish to know how well they did classifying the news articles, as it was anticlimactic to complete the task without receiving feedback on their accuracy. All of this feedback was incorporated to improve the final experiment design. For the experiment, the participants would see the number of correctly classified news articles in the end.

4.1.3. Stimulus Selection

The number of news articles was reduced for the experiment to improve the concentration of the participants throughout the session. Participants from the pilot test suggested using between 10 and 20 articles in total. Consequently, 4 out of 10 news articles were chosen per category as the final stimuli for the experiment, totalling 16 final news articles. As participants were not able to easily labelled human or AI-generated news from the pilot study, we decided to select the more easily identified, less misleading news as the final stimulus. This was done by assigning a score value to each news article. A news article gets one point for every participant that correctly classifies it as true or fake news, and another point if it is correctly classified as either human or AI-generated. Therefore, the maximum score for one news article of 14 was achieved when all the participants from the pilot test (N = 7) correctly classified both its truthfulness (true / fake) and origin (generated by humans / AI). The four news articles per category with the highest score were selected, totalling 16 news articles as stimuli for the experiment.

Origin	Authenticity	
	True	Fake
Human	H-T (4)	H-F (4)
AI	AI-T (4)	AI-F (4)
# of news:	16	

Table 4.2: Study design: the number of news per category used as the final stimuli dataset.

4.2. Participants and Recruitment

The Human Research Ethics Committee (HREC) of Delft University of Technology approved the study (application number *3533*). The data management procedures were consulted with the data stewards of the TU Delft EEMCS faculty. The datasets of the study have been made available on *4TU.ResearchData*[1].

The participants for the main experiment were recruited through posters at the university and via social media groups. Participation was incentivized by offering volunteers a 10 euro voucher, regardless of whether they successfully finished the experiment or not. There were 35 people recruited for the study and all of them completed the experiment. Out of all the participants, 19 identified as male, 15 as female, and 1 as non-binary. The average age was 27.08 (SD = 7.48, range = 23 – 57). The education levels varied among participants. 20 participants have a bachelor's degree, 14 have a master's degree, and 1 does not have a degree. The participants had a wide range of ethnicity, 15 were Caucasian, another 13 were Asian, 3 were of mixed ethnicity (2 were Asian & Caucasian, 1 was Black & Caucasian), and 4 were of Latino or Hispanic background.

4.3. Procedure

Figure 4.2 illustrates the experiment procedure. Participants are first provided with the consent form and given sufficient time to read, process the information and ask questions. The consent form informs the participants what the goals and tasks of the ex-

[1] https://data.4tu.nl/datasets/89d49b4a-9965-4122-bfcf-9df5505bf21b

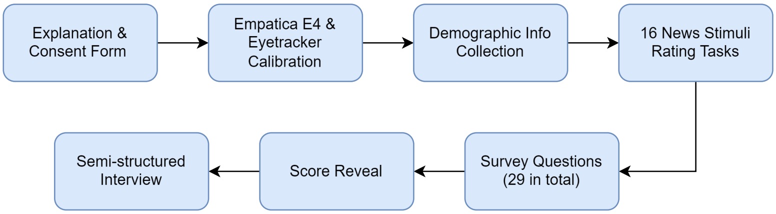

Explanation & Consent Form
Empatica E4 & Eyetracker Calibration
Demographic Info Collection
16 News Stimuli Rating Tasks
Survey Questions (29 in total)
Score Reveal
Semi-structured Interview

Figure 4.2: The figure demonstrates the experiment procedure in steps.

periment are. Some other information provided are the estimated time for completion, and details on what data is captured, as well as how the data is processed and stored. Participants are informed that they can opt out of the study at any given time.

The Empatica E4 sensor is placed on the wrist of the participant, then paired and connected via Bluetooth to the mobile app. We ask the participants to sit comfortably and adjust the height of the chair if needed. The participants are advised to minimize covering or touching their face during the experiment.

The calibration process of the Tobii Eye tracker requires participants to follow the dots on the screen with their gaze. Therefore, the calibration results are evaluated based on the captured gaze coordinates from the eye tracker versus the actual position of the dots on the screen. The experiment uses a 9-point calibration procedure. Tobii [63] defined good accuracy as all participants receiving average accuracy of less than 0.8°. Across all data points, the accuracy should not exceed 5°, while the precision standard deviation should not exceed 1.5°. It is suggested under ideal conditions, average precision must be $< 0.5°$ for good precisions. Across the 35 participants in the experiment, all of them met the criteria, with an overall average calibration accuracy of 0.26° (SD=0.12°, Max=0.59°, Min=0.11°) and average precision of 0.42° (SD=0.26°, Max=0.96°, Min=0.11°).

Successful calibration is followed by the Tobii Pro Lab environment, which leads participants to the PhysioNews app. The introduction screen provided the information about the experiment once again and stressed that participants should read the news carefully and read it at their own pace. After that, participants were asked to fill in information about their demographics.

After the introduction, participants begins reading the news and fill in their perception of truthfulness and origin of the news. News stimuli are proceeded with survey questions from different sources of questionnaire as described in section 3.8, and the full list of survey questions can be found in the Appendix.

The order of the stimulus follows a Latin-square design [64]. Based on the participation order, each person is assigned to a user group. The 4 user groups are *T-H*, *F-H*, *T-AI*, *F-AI*. The first stimulus shown to the user is controlled. The user group determines the category of the first stimulus shown to the participant. The order of the rest of the stimuli is random.

In the interview, participants are asked to share their experience with the experiment. In addition, questions were asked to better understand their decision process when rating the news. The following questions are asked:

- How was the overall experience of the experiment?

- (TRUE/FAKE) Were you able to easily tell the difference between true and fake news? why or why not? What strategies did you employ?
- (HUMAN-AI) Were you able to easily tell the difference between human and AI news? why or why not? What strategies did you employ?
- (DOMAIN) were there some types of news, e.g., topic, that you felt were easier or more difficult? why or why not?
- Do you use Chat-GPT regularly, or have you used it before? Do you think that it helped you with identifying AI-generated news in this experiment?
- Have you encountered such types of true-fake, human-AI news in your daily life? How did you deal with that? Where did you see it?

Follow-up questions were then asked based on the individual participant's answers.

4.4. DATA PREPROCESSING

Data from Tobii and Empatica were synchronized using Universal Time Coordinated (UTC) timestamps. Tobii Pro Lab categorized each domain visit within the experiment as a Time of Interest (TOI) interval, which gives the start and end times of each stimulus. For more precise analysis of eye movement visits on the news area, we defined Areas of Interest (AOI). For each news stimuli to differentiate the eye visits within and the news text are. Using AOI areas allowed us to more accurately evaluate the domain visits for each TOI. Additionally, the Savitzky-Golay filter is applied to smooth the pupil diameter measures. This filter was set with a polynomial order of 2 and a window length of 15 as suggested in the previous work [6].

For the Empatica data, we processed the raw BVP (blood volume pulse) data using the Neurokit Python library [65]. Baseline heart rate values were derived from the period when participants were filling out the demographic information form, occurred before the first stimulus was presented. We then computed heart rate features (both EDA and HRV features), as the ratio of the mean value during each TOI to the baseline, which is the mean value captured during the demographic screen. The ratios are used as the normalized measures of heart rate signals across all participants on the stimuli.

4.4.1. TOBII I-VT FILTER

We used the built-in Tobii I-VT (Fixation) filter from Tobii Pro Lab to process the raw data. The I-VT (algorithm 1) classifier has a threshold of 30 degree per second, which entails that gaze movement slower than 30° per second are categorized as fixations, while faster movements are classified as saccades. Additional parameters were set to eliminate short fixations and to merge adjacent fixations.

4.4.2. REGRESSIVE SACCADES

We adopted a similar approach to Bozkir et al. [37] in further classifying saccades within AOIs as regressive or progressive. Regressive saccades are identified under the following conditions when reading the news stimuli in the experiment:

1. The saccade moves in the negative direction of the x-axis while the gaze on the y-axis remains approximately on the same line in the news article. This indicates that the reader is going back on the same line. Given that the height of one line of text is approximately 20 pixels, we define this condition as the change in the x-axis being negative, but not exceeding the width of a line. This is characterized by the gaze moving to the left (negative x) while staying within a vertical range of ± 20 pixels on the y-axis.

2. The second condition involves any gaze moving in the positive direction on the y-axis that is larger than the height of a line, i.e. the gaze travels more than 20 pixels on the y-axis. We classify such saccades as regressive.

4.5. Technical Validation

Experiment procedures from Tobii[63] suggested that data is considered valid when at least 80% of collected gaze data is valid. They also defined that validity of data is when both the left and right eyes were captured and have validity as valid. The same measure is used in the research conducted by Sumer[6]. The data collected from a single stimulus of a participant is considered valid when 80% of the collected data during the TOI is valid. We excluded 9 participants out of 35 participants due to not enough valid data. The invalidity of the data is due to the participants getting too close to the screen, which caused the eye tracker the failure in capturing the eye(s). Even though they appeared to have good accuracy in calibration to begin with, they got too close to the screen during the session. For the analysis of HR related data, 2 participants were excluded due to Empatica device connection failure. In total, 33 participants data were used in the analysis of HR+EDA features.

4.6. Machine Learning Classifiers

Following a similar approach to previous work [7], we explore two types of classifiers: a user-independent classifier and a user-dependent classifier. A user-independent classifier aims to generalize across all users, with accuracy determined by the mean of 10-fold cross-validation. In contrast, a user-dependent classifier trains data separately for each user. The accuracy of the user-dependent classifier is calculated as the mean accuracy across all individual user-dependent classifiers. For each user-dependent model, the accuracy is derived from the mean of 4-fold cross-validation. We train four types of classifiers to predict: 1) whether the news is fake 2) whether the news is AI-generated 3) whether the reader perceives it as fake news, and 4) whether the reader perceives it as AI-generated news. As baseline models, we used Logistic Regression, Support Vector Machines, and Random Forest. All classifiers are set with `class_weight='balanced'` to better handle the imbalanced data distributions. Additionally, `random_state=42` is used for the SVM and RF classifiers to ensure replicability.

Four feature sets were used for this study, as shown in table 4.3. The first three sets (Gaze only, HR only, and Gaze + HR) attempt to train classification models purely from physiological signals from the participants. For the fourth feature set, in addition to the Gaze and HR features, we added more information about the specific news stimuli. If the task involves "AI news" or "perceived AI", then information on its true or fake label is incorporated, along with whether the participant perceived it as true or not and their rating of its truthfulness and familiarity with the news. Conversely, for the task that is "Fake news" or "perceived fake", additional information related to whether it is AI-generated or not is added to the feature set.

For all datasets, the NewsId is removed to prevent overfitting, as the True/Fake, Human/AI labels are directly linked to the NewsId. Meanwhile, the ParticipantId is incorporated, as it helps to recognize each individual user's attributes and tendencies.

Number	Feature set	Number of features
1	Gaze only	9
2	HR only	9
3	Gaze + HR	17
4	All (Gaze + HR + news context)	21

Table 4.3: The four feature sets used for the classification models.

5

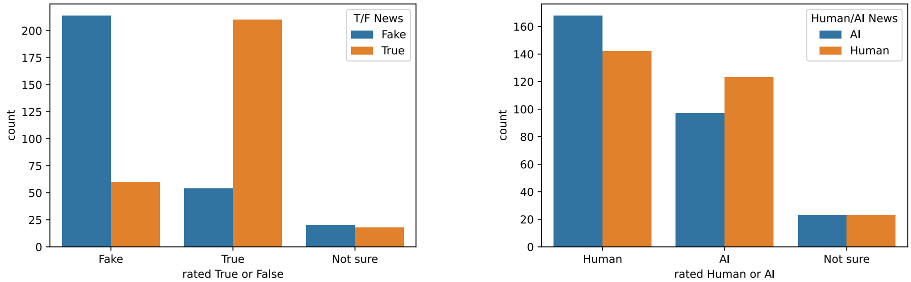
RESULTS

In this chapter, we present the results of the study. The results include the human perception of news based on participants' ratings (RQ1), statistical tests on physiological response features, and the performance of machine learning models using these features (RQ2). Lastly, we include statistical tests to determine if personality traits gathered from the questionnaire correlate with participants' performance in classifying the news (RQ3). After presenting the data of the results, we summarize the findings of the physiological response features in section 5.5 n as we interpret the results. These findings help us answer the research questions in the following chapter.

When presenting the results, the asterisk sign indicates the level of significance of the p-value. "***" means p<0.001, "**" indicates p<0.01, "*" means p<0.05 and "." indicates that p<0.1.

5.1. HUMAN PERCEPTION TOWARDS NEWS

Table 5.1 shows how the news was labeled among 35 participants (560 rows in total). The two tasks were the classification of true/fake news and human/AI-generated news. Based on the responses, the results show that the participants were relatively good at distinguishing true from fake news, with an accuracy of 72.9% for true news and 74.3% for fake news. However, their performance declined when classifying human vs. AI-generated news, with only 49.3% of human-generated news correctly identified and merely 33.6% of AI-generated news correctly classified. Notably, a considerable number of human-generated news items were misclassified as AI-generated, and vice versa. These findings are visualized in figure 5.1.



(a) True or Fake News count

(b) Human or AI-generated News count

Figure 5.1: Counts of news classification (N=35, row=560)

	labelled as			
News Type	True	Fake	Human	AI
True	**72.9% (210)**	20.8% (60)	64.9% (187)	28.4% (82)
Fake	18.7% (54)	**74.3% (214)**	42.7% (123)	47.9% (138)
	Human	AI	True	Fake
Human	**49.3% (142)**	42.7% (123)	40.6% (117)	54.5% (157)
AI	58.3% (168)	**33.6% (97)**	51.0% (147)	40.0% (117)

Table 5.1: Distribution of labeled news: Main experiment

Further analysis is shown in figure 5.2, which breaks down mean news ratings. According to figure 5.2a, true news received similar truthfulness ratings regardless of whether they were AI or human-generated. In contrast, AI-generated fake news was consistently rated as more convincing than its human-generated counterparts. This indicates that AI-generated content can appear to be more plausible than Human-generated ones.

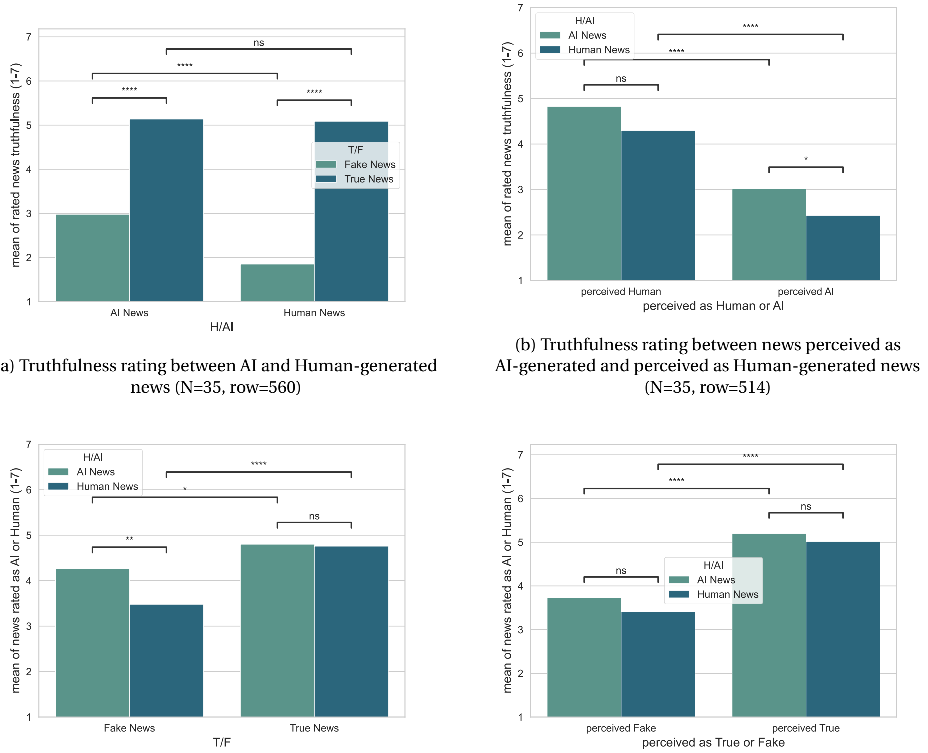
Figure 5.2b explores the mean truthfulness rating when news was perceived as human or AI. In both "perceived as human" and "perceived as AI" categories, AI-generated news received higher ratings. Generally, news perceived as human received significantly higher ratings in truthfulness than news perceived as AI.

From figure 5.2c, we observe that true news was rated similarly in terms of perception of whether it is human-generated or not, regardless of the actual source. However, AI-generated fake news was considered more human-like than fake news created by humans.

Figure 5.2d shows that in both "perceived as fake" and "perceived as true" categories, AI news received slightly higher ratings for being human-generated. Overall, news perceived as true received significantly higher ratings for being human-generated.

5.2. PHYSIOLOGICAL RESPONSES TO NEWS

The gaze features and the HR / EDA features (from E4) were captured. Due to different validation criteria (Section 4.5), the number of valid participants varied between the datasets: 26 participants for the Gaze feature dataset and 33 participants for the HR/EDA



(a) Truthfulness rating between AI and Human-generated news (N=35, row=560)

(b) Truthfulness rating between news perceived as AI-generated and perceived as Human-generated news (N=35, row=514)

(c) Mean rating of Human or AI-generated news scale in Fake and True news (N=35, row=560)

(d) Mean rating of Human or AI-generated news scale in news perceived as Fake and perceived as True

Figure 5.2: Average ratings of different news categories (N=35, row=525)

Dataset	N	Response Variable	W	P
All responses	35	Rating Fake or True (1-7)	0.86656	<0.0000
		Rating AI or Human (1-7)	0.9055	<0.0000
Gaze	26	Rating Fake or True (1-7)	0.85975	<0.0000
		Rating AI or Human (1-7)	0.89547	<0.0000
		Duration (sec)	0.95885	<0.0000
		Saccade count	0.99143	0.01945
		Fixation count	0.97988	<0.0000
		Pupil diameter mean	0.98067	<0.0000
Heart Rate	33	Rating Fake or True (1-7)	0.86841	<0.0000
		Rating AI or Human (1-7)	0.90736	<0.0000
		HRV RMSSD	0.27112	<0.0000
		HRV MeanNN	0.58958	<0.0000
		HRV SDNN	0.3209	<0.0000
		HRV SDSD	0.26991	<0.0000
		BPM	0.78172	<0.0000
		Clean EDA	0.49183	<0.0000
		Tonic EDA	0.47968	<0.0000
		Phasic EDA	0.37419	<0.0000

Table 5.2: Results of Shapiro-Wilk normality tests

features. Each feature set was analyzed separately using ANOVA and the Mann-Whitney U test. Later, these sets were combined for the training of ML classifiers, with a subset of 24 participants (Section 5.3).

5.2.1. NORMALITY TEST

The Shapiro-Wilk normality test was used to check the data distribution and determine whether the assumption of normality is met for other statistics tests. Table 5.2 showed that all features, including gaze and heart rate features, all significantly deviated from a normal distribution. To proceed with ANOVA, which assumes normality, we applied aligned rank transforms (ART). ART makes the data meet the distribution requirements with a non-parametric technique, which aligns and ranks data before analysis [66]. This adaptation ensures that the assumptions of ANOVA are met.

5.2.2. ANOVA ANALYSIS

GAZE FEATURE

Table 5.3 presents the of analysis of variance for selected gaze features. The dataset includes 26 participants. The table is split into two sections: results of the first part are based on factors of actual label (TF, HAI), while the second part concerns factors from perception of the user (pTF, pHAI). Statistically significant differences were observed in several variables, particularly in saccade count and fixation count with the H/AI factor (F=140 for saccade count, F=187 for fixation count). Noticeably, fewer statistically significant differences and lower F-values were observed in tests on grouped by perception of the news (pTF, pHAI).

	Variable	Factor	F	df	p	
Rating	T/F rating	TF	177.4709	1	<0.000	***
		HAI	11.467	1	<0.000	***
		TF × HAI	9.8774	1	<0.00	**
	H/AI rating	TF	24.5645	1	<0.000	***
		HAI	2.1476	1	0.1436	
		TF × HAI	3.4708	1	0.0632	.
duration	duration in AOI	TF	15.2606	1	<0.000	***
		HAI	8.4615	1	<0.00	**
		TF × HAI	11.0581	1	<0.000	***
saccade	saccade count	TF	51.44	1	<0.000	***
		HAI	**140.885**	1	<0.000	***
		TF × HAI	21.041	1	<0.000	***
fixation	fixation duration mean	TF	2.66951	1	0.10313	
		HAI	0.36281	1	0.54732	
		TF × HAI	22.92742	1	<0.000	***
	fixation count	TF	75.556	1	<0.000	***
		HAI	**187.318**	1	<0.000	***
		TF × HAI	11.415	1	<0.000	***
pupil	mean pupil diameter	TF	6.095884	1	0.013995	*
		HAI	0.062264	1	0.803089	
		TF × HAI	1.081807	1	0.298964	
	mean pupil diameter fixation	TF	6.234956	1	0.012953	*
		HAI	0.067925	1	0.794526	
		TF × HAI	1.078881	1	0.299617	
duration	duration in AOI	pTF	5.721963	1	0.017241	*
		pHAI	0.006412	1	0.936221	
		pTF × pHAI	1.175985	1	0.278866	
saccade	saccade count	pTF	4.52439	1	0.034058	*
		pHAI	1.55712	1	0.212854	
		pTF × pHAI	0.72582	1	0.394781	
fixation	fixation duration mean	pTF	1.0726	1	0.30102	
		pHAI	0.94472	1	0.33169	
		pTF × pHAI	0.29908	1	0.58478	
	fixation count	pTF	3.69604	1	0.055284	.
		pHAI	1.16333	1	0.281455	
		pTF × pHAI	0.60662	1	0.436546	
pupil	mean pupil diameter	pTF	12.06648	1	<0.000	***
		pHAI	0.91711	1	0.338851	
		pTF × pHAI	0.38036	1	0.537786	
	mean pupil diameter fixation	pTF	12.11838	1	<0.000	***
		pHAI	0.89935	1	0.343566	
		pTF × pHAI	0.40091	1	0.527006	

Table 5.3: Results of ANOVA, conducted on Gaze features processed with ART

HR FEATURES

From table 5.4, we gathered the ANOVA analysis results of HR features using the dataset with 33 participants. The HR variables with statistical significance are predominantly from the factors of user perception (pTF, pHAI). RMSSD in the factor of pTF (perceived true or fake) showed the highest F-value of 31.11. Tonic EDA also showed some levels of significance in pTF and pHAI. Generally, the effect size is much smaller than the gaze features from table 5.3. ndtable

Figures 5.6a, 5.6b, 5.6c and 5.6d are gaze features on actual news labels; figures 5.3e, 5.3f, 5.3g and 5.3h are HR features on news perception labels.

5.2.3. MANN-WHITNEY

Figures 5.6a, 5.6b, 5.6c and 5.6d displays the normalized gaze features visualized across the four news categories using violin plots for "Human-True" (H-T), "Human-Fake" (H-F), "AI-True" (AI-T), and "AI-Fake" (AI-F). Figures 5.3e, 5.3f, 5.3g and 5.3h illustrated the HR features on news perception labels. More results can be found in Appendix. Each plot shows the distribution and median value. Statistical analysis was conducted with the Mann-Whitney test to compare each pair of categories. The results showed significant differences between multiple pairs for saccade count and fixation count, while no significance found for HR features. The "*" sign indicates levels of statistical significance.

5.3. MODELING NEWS LABELS & USER PERCEPTION OF NEWS

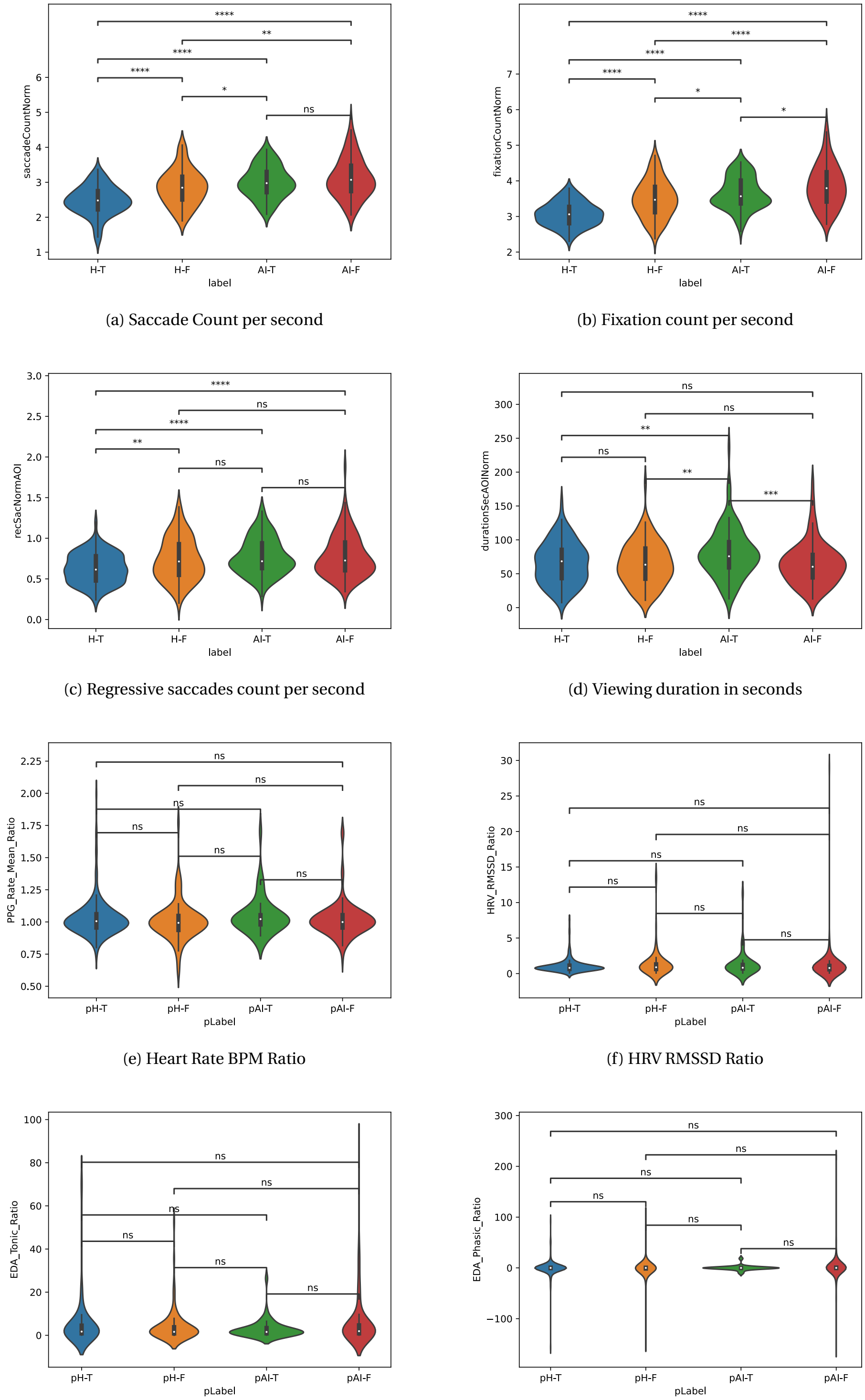
The feature sets and classification tasks are explained in section 4.6. The four tasks are: predicting if 1) the news is fake 2) the news is AI-generated 3) reader perceived it as fake news 4) reader perceived it as AI-generated news. Each model is then trained with Linear Regression (LR), Support Vector Machines (SVM), and Random Forrest (RF) classifiers. The machine learning models were all trained on the dataset with 24 participants, with 368 number of rows.

5.3.1. USER-INDEPENDENT - BASELINE

Figure 5.4 shows the results of the baseline models using of four different feature sets (Gaze only, HR only, Gaze + HR, All), with three ML classifiers (LR, SVM, RF), and the four different prediction tasks. The result of each model is the average score of 10-fold cross validation. The values highlighted in bold represent the best performing model for each task. Using the combination of all three feature sets yields the best results in predicting fake news, perceived fake, and perceived AI. However, using only the gaze feature set yields the best results in predicting AI news at 78.36%.

	Variable	Factor	F	df	p	
Rating	T/F rating	TF	280.079	1	<0.0000	***
		HAI	20.823	1	<0.0000	***
		TF × HAI	13.612	1	0.00025	***
	H/AI rating	TF	35.2561	1	<0.0000	***
		HAI	8.7245	1	0.003291	**
		TF × HAI	8.7138	1	0.00331	**
Heart Rate	HRV RMSSD	TF	0.22876	1	0.63266	
		HAI	3.16424	1	0.07589	.
		TF × HAI	1.4544	1	0.22841	
	HR BPM Mean	TF	0.66601	1	0.414844	
		HAI	4.74596	1	0.029845	*
		TF × HAI	0.46042	1	0.497749	
	HRV MeanNN	TF	0.15984	1	0.6894793	
		HAI	7.61450	1	0.0060077	**
		TF × HAI	0.76109	1	0.3834165	
EDA	EDA Tonic	TF	3.4665	1	0.063225	.
		HAI	0.3714	1	0.542525	
		TF × HAI	4.3707	1	0.037078	*
	EDA Phasic	TF	0.034991	1	0.85169	
		HAI	0.285458	1	0.59339	
		TF × HAI	0.047697	1	0.82721	
	EDA Clean	TF	3.76142	1	0.053024	.
		HAI	0.43181	1	0.511414	
		TF × HAI	4.56324	1	0.033162	*
Heart Rate	HRV RMSSD	pTF	31.11	1	<0.0000	***
		pHAI	10.9075	1	0.001027	**
		pTF × pHAI	1.5295	1	0.216777	
	HR BPM Mean	pTF	0.078938	1	0.77886	
		pHAI	6.649693	1	0.010207	*
		pTF × pHAI	1.282625	1	0.257965	
	HRV MeanNN	pTF	6.07924	1	0.014018	*
		pHAI	0.48319	1	0.487309	
		pTF × pHAI	1.20240	1	0.273380	
EDA	EDA Tonic	pTF	12.4741	1	0.000451	***
		pHAI	9.6865	1	0.001963	**
		pTF × pHAI	0.352	1	0.553257	
	EDA Phasic	pTF	1.29258	1	0.2561	
		pHAI	0.044901	1	0.83227	
		pTF × pHAI	2.418951	1	0.12048	
	EDA Clean	pTF	10.2018	1	0.0014922	**
		pHAI	8.9875	1	0.0028540	**
		pTF × pHAI	0.5208	1	0.4708405	

Table 5.4: Results of ANOVA, conducted on HR features processed with ART



(a) Saccade Count per second

(b) Fixation count per second

(c) Regressive saccades count per second

(d) Viewing duration in seconds

(e) Heart Rate BPM Ratio

(f) HRV RMSSD Ratio

(g) EDA Tonic Ratio

(h) EDA Phasic Ratio

Figure 5.3: Violin plots of features. Figures 5.6a, 5.6b, 5.6c and 5.6d are gaze features on actual news labels; figures 5.3e, 5.3f, 5.3g and 5.3h are HR features on news perception labels.

Feature	Mode	Fake news		AI news		perceived fake		perceived AI	
		acc	F1	acc	F1	acc	F1	acc	F1
Gaze only	LR	58.47%	56.35%	70.73%	69.11%	48.08%	45.92%	45.11%	39.08%
	SVM	57.91%	56.24%	**78.36%**	74.32%	49.15%	43.14%	45.18%	42.81%
	RF	56.54%	56.78%	73.73%	68.59%	51.04%	48.05%	59.50%	39.59%
HR only	LR	53.55%	49.54%	49.42%	39.72%	53.00%	48.49%	48.64%	41.31%
	SVM	48.93%	35.30%	50.52%	26.94%	51.09%	32.48%	50.26%	36.72%
	RF	49.45%	49.67%	51.66%	48.95%	52.22%	49.53%	55.13%	34.91%
Gaze + HR	LR	59.26%	58.26%	70.13%	67.96%	47.52%	44.64%	45.94%	38.57%
	SVM	57.37%	56.60%	78.36%	74.20%	51.62%	43.48%	51.39%	42.22%
	RF	56.01%	56.13%	70.21%	65.70%	47.83%	42.81%	59.81%	35.28%
All	LR	65.53%	66.02%	72.32%	70.50%	64.95%	**64.89%**	66.88%	62.07%
	SVM	**66.91%**	**69.77%**	77.25%	**75.41%**	**65.56%**	61.54%	**69.35%**	**66.45%**
	RF	59.00%	60.69%	71.01%	66.63%	66.64%	64.35%	66.07%	53.63%

Figure 5.4: User-Independent baseline ML classifiers results, with 10-fold cross validation

5.3.2. User-Independent - PCA

Principal Component Analysis (PCA) is a technique commonly used for dimensionality reduction that often improves performance of machine learning models on tabular data [67]–[69]. The original values of the features are transformed into principal components, which retain the information of the original variables. Components are ordered by the information they hold. For our study, for each feature set, the number of principal components used as new input features is the number that accounts for at least 80% of explained variance.

The PCA scree plots and performance metrics provided insights into the effectiveness and characteristics of the gaze and heart rate feature sets. The PCA scree plots and performance chats presented in figure 5.5 were on the task of predicting AI news with all four different feature sets. For other tasks (Fake news, perceived as fake, or perceived as AI), the results can be found in the More results can be found in Appendix.

Table 5.5 includes the results with PCA applied, using the number of components (nc) that captured at least 80% of the variance. Figure 5.5 shows the progression of cumulative explained variance. The threshold of 80% is reached at nc = 4 for Gaze feature set, nc = 2 for Heart rate feature set; nc= 5 for Gaze and HR feature sets, and also nc = 5 for the All feature set. In table 5.5, the accuracy and F1 values are highlighted in bold when they are higher than the baseline classifier. From figure 5.5, we also observed higher accuracy reached when more number of components were incorporated with the Gaze only, Gaze + HR and All feature sets.

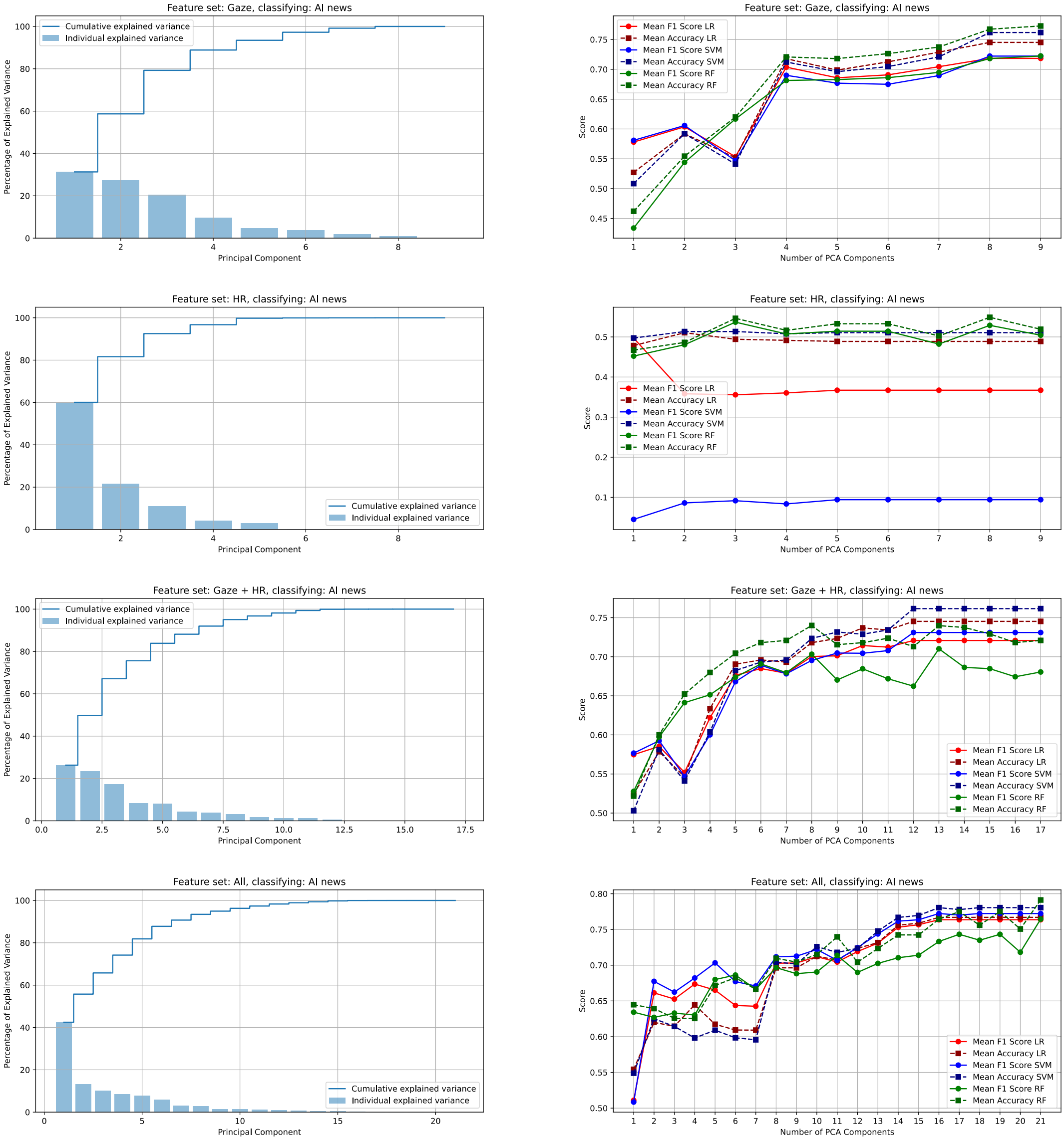


Figure 5.5: PCA analysis of predicting if news is AI-generated or not.

Feature	Mode	Fake news		AI news		perceived fake		perceived AI	
		acc	F1	acc	F1	acc	F1	acc	F1
Gaze only	LR	**59.20%**	56.35%	**73.09%**	**71.53%**	51.65%	42.06%	54.17%	34.17%
	SVM	58.59%	**57.53%**	69.01%	65.35%	51.04%	39.76%	54.17%	33.85%
	RF	55.30%	52.81%	70.14%	64.24%	**59.64%**	43.72%	59.11%	29.19%
HR only	LR	47.14%	44.62%	55.56%	49.90%	42.53%	36.68%	52.08%	36.42%
	SVM	47.05%	42.36%	55.73%	46.32%	45.49%	37.72%	51.04%	37.60%
	RF	48.96%	50.24%	50.52%	42.99%	50.17%	36.10%	57.81%	31.59%
Gaze + HR	LR	55.21%	53.33%	70.05%	65.42%	45.92%	36.04%	56.25%	35.13%
	SVM	55.99%	54.17%	69.97%	65.59%	47.48%	38.47%	53.47%	33.96%
	RF	52.26%	51.04%	67.62%	61.87%	54.34%	39.02%	61.98%	33.77%
All	LR	55.56%	53.54%	66.67%	63.47%	59.20%	48.87%	62.67%	42.29%
	SVM	54.60%	53.85%	65.10%	62.29%	59.29%	**50.32%**	61.89%	**42.59%**
	RF	52.43%	49.97%	67.62%	64.48%	59.46%	43.42%	**66.75%**	37.66%

Table 5.6: Results of User-Dependent classifiers, with 4-fold cross validation mean accuracy of different feature sets; highest accuracy per prediction category in bold

Feature	nc	Mode	Fake news		AI news		perceived fake		perceived AI	
			acc	F1	acc	F1	acc	F1	acc	F1
Gaze only	4	LR	**59.54%**	**58.97%**	71.79%	**70.37%**	**52.70%**	**49.82%**	44.58%	35.28%
		SVM	**60.59%**	**63.02%**	71.23%	69.00%	51.06%	41.19%	46.20%	32.56%
		RF	54.09%	54.83%	72.08%	68.14%	48.32%	45.92%	54.35%	32.98%
HR only	2	LR	52.45%	41.16%	**51.06%**	**35.82%**	52.44%	45.98%	**54.04%**	**44.43%**
		SVM	49.98%	6.74%	51.34%	8.60%	49.45%	34.31%	52.42%	38.38%
		RF	**53.56%**	54.50%	48.63%	48.06%	52.74%	49.73%	54.94%	37.86%
Gaze + HR	5	LR	59.26%	58.56%	69.05%	67.62%	**52.42%**	**49.22%**	45.13%	33.87%
		SVM	**59.79%**	**62.60%**	68.24%	66.81%	52.70%	44.48%	45.92%	29.28%
		RF	**58.70%**	**59.26%**	**70.45%**	**67.38%**	**54.56%**	**50.22%**	**62.74%**	**46.13%**
All	5	LR	58.21%	59.42%	61.73%	66.51%	**67.96%**	64.35%	68.81%	**65.50%**
		SVM	61.47%	65.17%	60.90%	70.33%	69.35%	65.37%	69.35%	66.45%
		RF	58.18%	58.09%	67.17%	**67.99%**	63.88%	60.85%	63.93%	54.04%

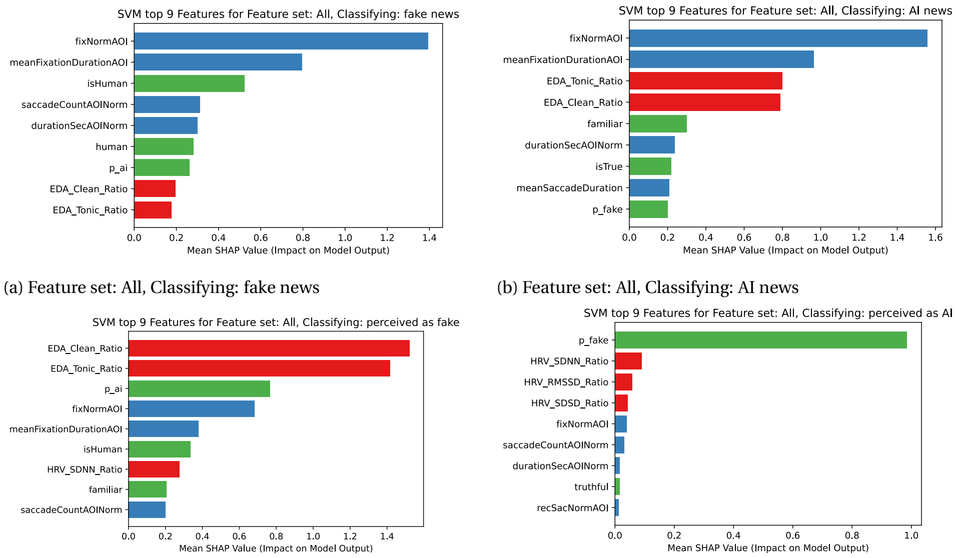
Table 5.5: 10-fold cross-validation mean accuracy and F1 score of different feature sets; accuracy and F1 scores higher than baseline model are highlighted in bold.

5.3.3. User-Dependent

From table 5.6, we observed that the classification accuracy of user-dependent models is generally lower than that of user-independent classifiers. Specifically, most prediction tasks performed best with the Gaze only feature set, in tasks involving Fake news, AI news and perceived fake news. This trend was visible despite that each prediction model was trained on data from only one user, with a maximum of 16 entries.

5.3.4. Feature Importance

To gain a better understanding of how features contribute to the model, we conducted SHAP analysis. Figure 5.6 presents the feature importance analysis using SHAP [70]. The plots illustrate the proportion of features contributing to the predictions. Only the results where all feature sets were enabled are shown here, to provide a comprehensive overview of the contributions of all features. Results of SHAP analyses for other feature sets and classification tasks are available in the Appendix.



(a) Feature set: All, Classifying: fake news

(b) Feature set: All, Classifying: AI news

(c) Feature set: All, Classifying: perceived as fake

(d) Feature set: All, Classifying: perceived as AI

Figure 5.6: SHAP Values - Impact on SVM Model Output, with all feature sets enabled.

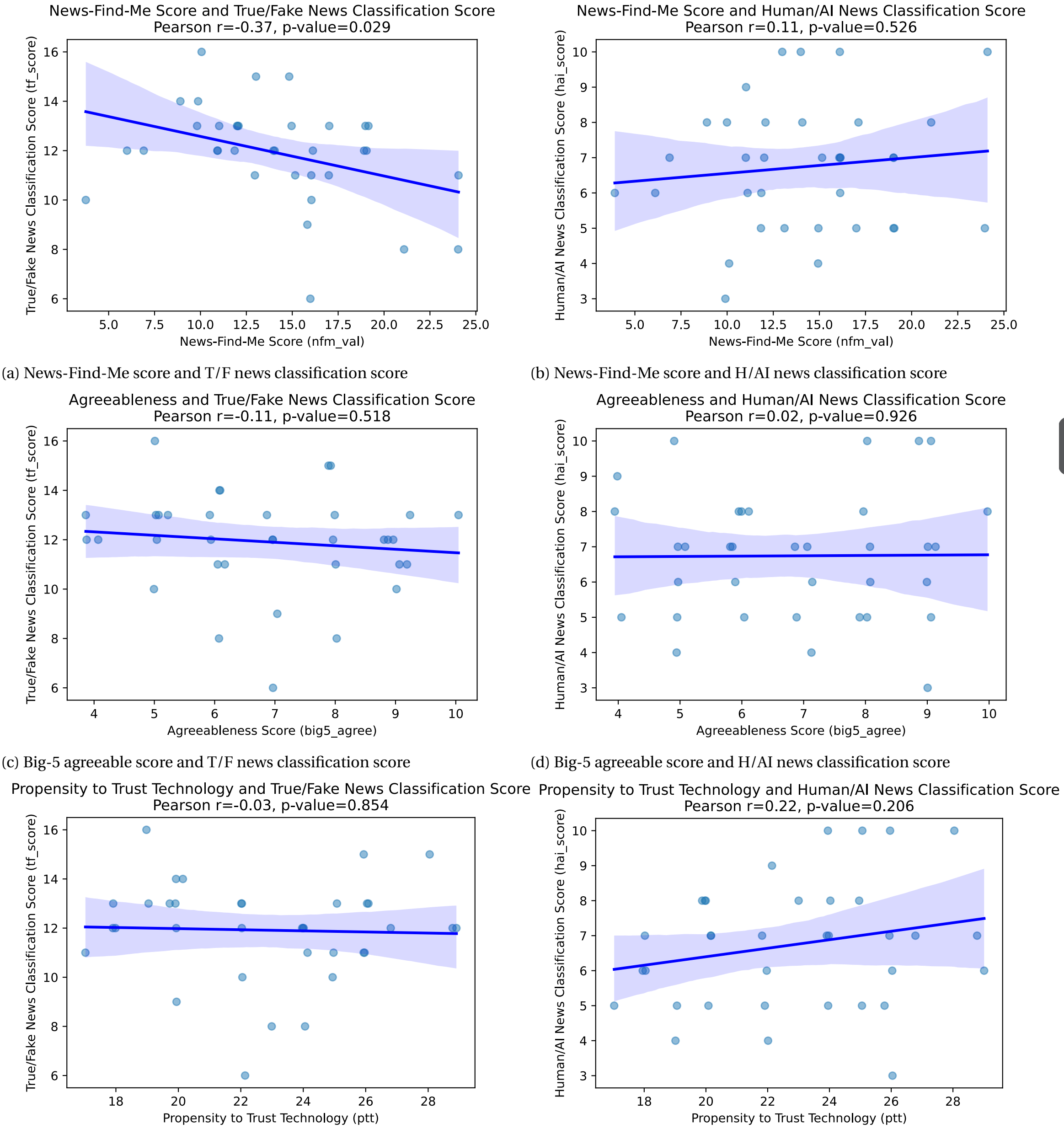
5.4. User's traits and News Classification Score Correlations

Figure 5.7 illustrates the relationship of T/F and H/AI news classification score of users and their trait. The three user traits we examined are News-Find-Me, Agreeableness from the Big-5 personality test, and Propensity to Trust in Technology, as described in section 3.8.

From figure 5.7a, we observe that users with lower News-Find-Me scores tend to score higher in classifying True and Fake news. This relationship is statistical significant with the Pearson correlation coefficient test (r=-0.37, p=0.029). In contrast, the relationship for Human/AI news classification, as shown in figure 5.7b, exhibits a slight tilt in the line but is statistically insignificant (r=-0.11, p=0.526).

Figures 5.7c and 5.7d show the relationships between the Big-5 Agreeableness trait and news classification. The correlation lines for both T/F and H/AI scores are flat and statistically insignificant.

Regarding the results of Propensity to Trust in Technology, shown in figures 5.7e and 5.7f, there is a noticeable trend where higher H/AI news classification scores occur when Propensity to Trust is higher. However, this trend is also statistically insignificant (r=0.22, p=0.206).



(a) News-Find-Me score and T/F news classification score

(b) News-Find-Me score and H/AI news classification score

(c) Big-5 agreeable score and T/F news classification score

(d) Big-5 agreeable score and H/AI news classification score

(e) Propensity to Trust score and T/F news classification score

(f) Propensity to Trust score and H/AI news classification score

Figure 5.7: Relationship and Pearson correlation test of News-Find-Me, Agreeable personality, and Propensity to Trust and T/F, H/AI classification scores.

5.5. Summary

5.5.1. ANOVA

From the ANOVA test presented in section 5.2.2, the results indicated that the gaze features correlate more strongly with the actual authenticity or origin of the news rather than what the participants perceive it as. This finding underscores the relevance of gaze feature in classifying the origin of news, whether human or AI-generated.

Our analysis from results, in machine learning classifiers and SHAP analysis indicates that gaze feature correlate more strongly with the actual authenticity or origin of the news rather than what the participants believe it to be.

Our analysis using ANOVA and the Mann-Whitney test showed that users displayed significantly more fixations and saccades while reading AI-generated news. In comparison, these differences were not reflected in the participants' perception of the news' authenticity or origin. The phenomenon of higher fixation and saccade counts in response to fake news, is likely due to the cognitive load associated with the nature of the content, as suggested by the literature [71], [72]. AI-generated content was suggested to require more cognitive processing than human written ones [26], which may be why more saccades were found in AI-generated news.

5.5.2. Classification and performance

From the results of the machine learning classifiers, when comparing the four feature sets and four prediction tasks, we observed that the AI news prediction task performs significantly better than other tasks. The gaze feature set achieved much better results than with the HR feature set. Although incorporating both feature sets and adding the True/ Fake news label (set 4) yielded similar results, the gaze feature set with SVM model scored the highest (78.36% accuracy, 74.32% F1). For predicting "Fake news", "perceived fake", and "perceived AI", using the All dataset achieved the highest score.

PCA

The gaze feature set and the heart rate feature set yielded vastly different results. For the gaze feature set, the performance increased as more PCA components were included. This suggests that even components with lower variance may hold useful information for classification. On the other hand, the scree plot for the heart rate feature set showed lower cumulative variance. The performance chart for this set remained relatively unchanged, while both the accuracy and F1 scores were considerably low. This could indicate that heart rate data contain less relevant information for the task of classifying AI news. For the Gaze + HR set and All set, most classifiers' performance maximized at a certain number of components and then plateaued. This shows that PCA can be an effective way to reduce the complexity of the model while maintaining good performance.

Performance of User-Dependent Classifiers

The relatively lower performance of the user-dependent model could stem from several factors. Primarily, the lack of data per user limits the model's training effectiveness. Using cross-validation further reduces the data available in the training set. A notable observation is that All feature set was less effective than the Gaze only feature set. This might be due to the lack of contextual information on how users' perception of news influenced their ratings compared to data aggregated from all participants. For example, as shown in figure 5.6, the variable p_fake (perceived as fake or not) significantly impacts the classification. The effect mainly derived from the dataset of all participants. In contrast, the data from a single user-dependent model are considerably more limited, which may affect the usefulness of All feature set.

5.5.3. SHAP

We noticed that, in general, models are more effective in predicting fake news or AI news than at predicting perceptions. From the SHAP analysis, we observed that the most influential feature in predicting whether news is perceived as AI is *p_fake*, which indicates whether the user thinks the news is fake. This finding shows a strong correlation between perceptions of news as AI and as fake. It reflects the prevalent bias among the participants in our study that associate fake news with AI-generated news. Additionally, fixation count is the highest ranked feature in terms of importance for predicting both fake and AI news.

6

Conclusion

Several findings were observed in chapter 5, which help us answer the three research questions. The research questions were about the perception of news (RQ1), the prediction of news labels based on the physiological features (RQ2), and the correlation between participants' scoring and their personality traits (RQ3).

RQ1: Do humans perceive human- and AI-generated news differently, and if so, can they correctly categorize the news?

The results suggest that while participants can generally distinguish true news from fake news, they struggle to classify the origin of the news (human vs. AI). Furthermore, the perceived truthfulness of the news is greatly influenced by whether it is believed to be human- or AI-generated. AI-generated news is often viewed as more plausible and human-like than actual human-generated news. These results align with findings from previous work, indicating that participants were more inclined to perceive AI-generated news as truthful compared to human-generated news [24]. In contrast, news that was believed to be AI-generated was rated lower in truthfulness. Similarly, news perceived as true was rated as significantly more human-like compared to news perceived as fake. This suggests that the perceptions of authenticity and origin are influenced by the nature of the news content.

RQ2: Can eye movements and heart rate data effectively predict users' perception of news' origin (Human or AI-generated) or authenticity (True or Fake)?

The gaze features – saccade count and fixation count highly correlate to the actual origin and authenticity of the news, as observed from the statistical tests using Mann-Whitney and ANOVA. However, there was little correlation between the gaze features and users' perception of the news' origin and authenticity. For our machine learning models, high performance was achieved in predicting whether news is AI or not using the Gaze feature set (Figure 5.4). From the SHAP analysis, we observed the importance of fixation count in the prediction models, since fixation count ranked highest in importance when predicting both Fake and AI news (Figure 5.6).

In comparison, heart rate data was much weaker in predicting the origin and authenticity of the news. Generally, predicting labels other than AI-generated news (Fake news, perceived fake, perceived AI) was less effective. When incorporating human perception (the "All" feature set), a higher performance of the model is achieved.

RQ3: Do personality traits, news-seeking behaviour, and trust in technology correlate with the susceptibility to falling for fake news?

While we did not find a significant correlation with the Big-5 agreeable personality and Propensity to Trust in Technology, it was shown that news seeking behavior correlates with the susceptibility to falling for fake news. The observation from the results is that users who are better at distinguishing true and fake news tend to score lower on New-Find-Me. In other words, scoring lower on News-Find-Me implies that they are more proactive in seeking news on their own. It could mean that they are more in tune with the news and current affairs, as they do not rely on other people to be informed. Therefore, they tend to do better in discerning True and Fake news.

6.1. Limitations

The participant count for our study was 35. The data of their ratings of news were used to understand their perception of news and observations were drawn. However, the dataset was reduced to 24 participants for the experiments that involved different feature sets, including gaze and heart rate data, due to data loss from the hardware of the eye tracker and heart rate sensor. Although data processing techniques were used to denoise and remove outliers, having a larger sample size would benefit the robustness of the study. Improved measurement practices could help minimize such data loss in the future.

For our study, we used the Tobii eye tracker and Empatica E4 for heart rate tracking. While these devices provided adequate results, the hardware precision and reliability of the physiological data were not optimal. This could impact the overall findings.

Selecting suitable news articles for the four categories (True/Fake and Human/AI-generated) posed challenges. Statistically, the number of words between the 4 categories could have been better matched to ensure consistency. Our final news dataset consisted of 16 news articles. Some participants felt that the news was not diverse enough. Originally, we intended to include 40 articles, but due to time constraints, only 16 were used in the final study after our pilot test.

6.2. Conclusion

In this thesis, we explored how humans perceive AI-generated news content and misinformation from a physiological perspective. We gathered and constructed a dataset of news with True/Fake, Human/AI-generated labels for our study. A web application was developed for the experimental purposes. The experimental setup is integrated with an eye tracker and a heart rate sensor. In total, 35 individuals participated in the study. We found that people struggle to distinguish AI-generated news from human-created news. In particular, gaze responses, specifically saccades and fixations, showed a high correlation with AI-generated misinformation, whereas correlations from heart rate data were less pronounced. Previous studies have shown that higher saccade and fixation count were found in Fake news compared to True news [6]. Our study found a similar discrepancy between Human and AI-generated news. The correlation between AI-generated news and higher saccade/fixation count is likely due to the higher cognitive processing load in AI-generated text, which induced more saccades and fixations in humans [71], [72]. Additionally, our investigation of users' personality traits and news-seeking behaviors revealed that individuals who scored higher on the "News-Find-Me" survey were better at discerning fake news and true news. We have made the experiment data, including gaze, heart rate, and news labeling, available for future research.

As AI advances, it is essential to thoroughly research how AI impacts humans. This is crucial to prevent Generative AI from being a means to an end for spreading misinformation. Our study highlighted that most people from the experiment cannot tell AI-generated content from non-AI-generated content apart. Therefore, spreading correct knowledge and raising awareness to the public about AI technology is vital. The general unfamiliarity with generative AI and the distrust among users are alarming and demonstrate that knowledge about AI, or the so-called "AI Literacy" [73] should be communicated more effectively. Legislation like the EU AI Act [74] is a step toward better informed users. Furthermore, our findings suggest that the significant increase in saccades and fixations when reading AI-generated content could help develop AI detectors. It can be especially useful for the emerging technologies of remote eye trackers, which monitor eye gaze via webcams [75], [76]. If integrated into a remote eye tracker, it could potentially help people identify AI-generated content.

Acknowledgements

I am extremely grateful for the opportunity to work with my supervisors, Pablo and Abdo. Their extraordinary support, encouragement, and guidance have been invaluable throughout my thesis research journey. I want to thank Pablo and Abdo for continuously inspiring me with new ideas and helping to refine my research.

I am thankful to be a part of the DIS group from CWI. I have learned a lot from sharing experiences and collaborating with the wonderful people I met there. Moreover, the atmosphere has always been very supportive and encouraging.

Special thanks to Ujwal from TU Delft for granting me access to the Tobii eye tracker from the very beginning of my research project. I am also thankful to CWI for providing the other equipment used for the study. Lastly, I am also grateful to all 35 people who took part in my studies. Without their involvement, this research would not have been possible.

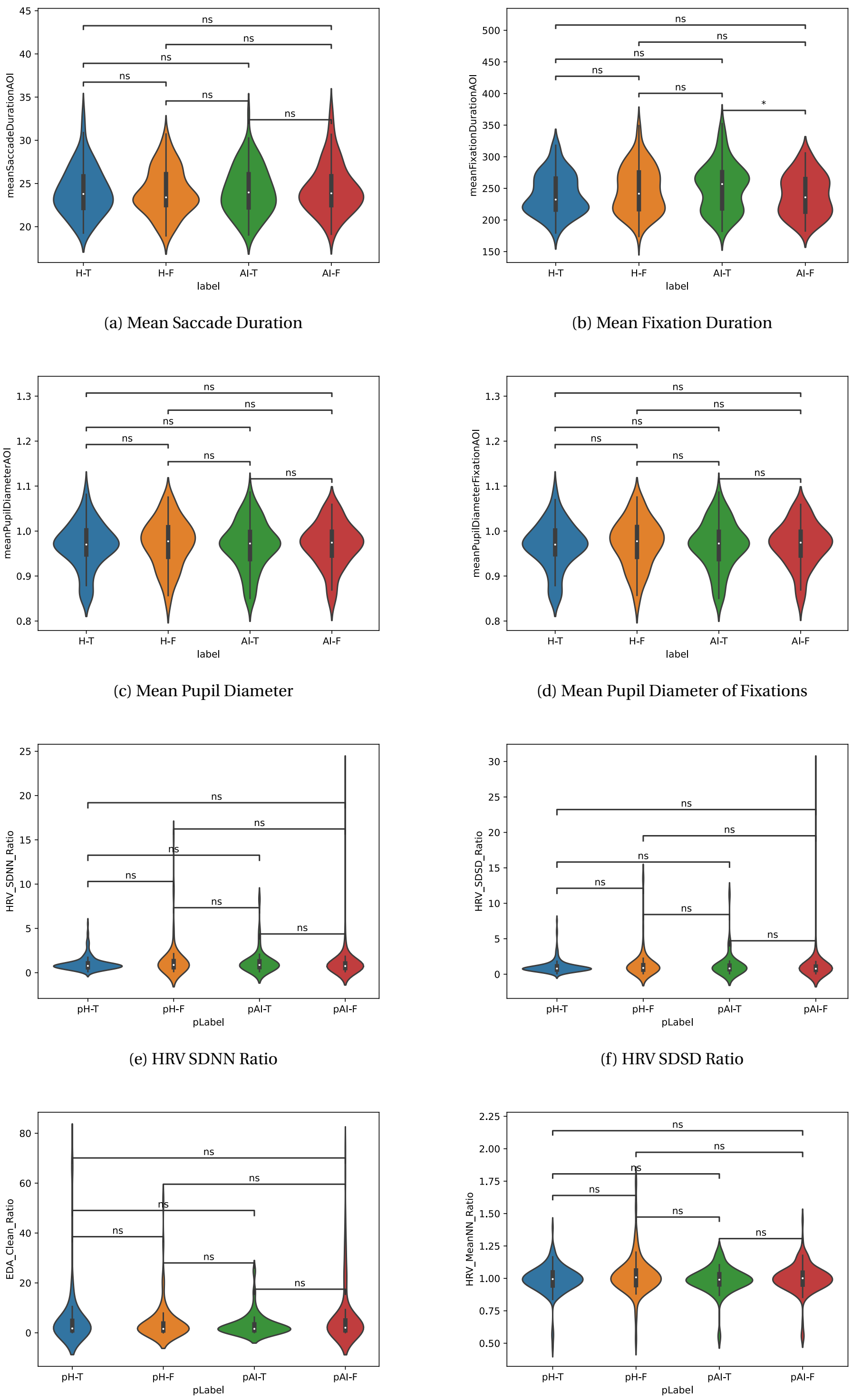
This experience has been profoundly valuable to me personally. Not only did I learn about how research is conducted, but I am also so thankful for the many encounters, experiences, and lessons that came with it. It has been about more than just the research itself.

Lian A. Wu
Den Haag, June 2024

APPENDIX

Item	Question	Likert scale
crt1	A soup and a salad cost a total of 5.50 euro. The soup costs one euro more than the salad. How much is the salad?	-
crt2	If 2 nurses take 2 minutes to measure the blood pressure of 2 patients. How long will it take for 200 nurses to measure the blood pressure of 200 patients?	-
crt3	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?	-
g1	How fluent are you in listening and reading in English?	7
g2	How fluent are you in writing in English?	7
g3	How often do you read / watch the news?	7
g4	How closely do you follow political news from the USA?	7
g5	How closely were you following the news on the Coronavirus pandemic since the outbreak in 2020?	7
nfm1	I rely on my friends to tell me what's important when news happens.	7
nfm2	I can be well-informed even when I don't actively follow the news.	7
nfm3	I don't worry about keeping up with the news because I know news will find me.	7
nfm4	I rely on information from my friends based on what they like or follow through social media.	7
po	When you think of your own political views, where would you classify your basic political stance?	7
ptt1	Generally, I trust technology.	5
ptt2	Technology helps me solve many problems.	5
ptt3	I think it is a good idea to rely on technology for help.	5
ptt4	I don't trust the information I get from technology.	5
ptt5	Technology is reliable.	5
ptt6	I rely on technology.	5
bfi1	I see myself as someone who is reserved	5
bfi2	I see myself as someone who is generally trusting	5
bfi3	I see myself as someone who tends to be lazy	5
bfi4	I see myself as someone who is relaxed, handles stress well	5
bfi5	I see myself as someone who has few artistic interests	5
bfi6	I see myself as someone who is outgoing, sociable	5
bfi7	I see myself as someone who tends to find fault with others	5
bfi8	I see myself as someone who does a thorough job	5
bfi9	I see myself as someone who gets nervous easily	5
bfi10	I see myself as someone who has an active imagination	5

Table 1: Questionnaire for the study



(a) Mean Saccade Duration

(b) Mean Fixation Duration

(c) Mean Pupil Diameter

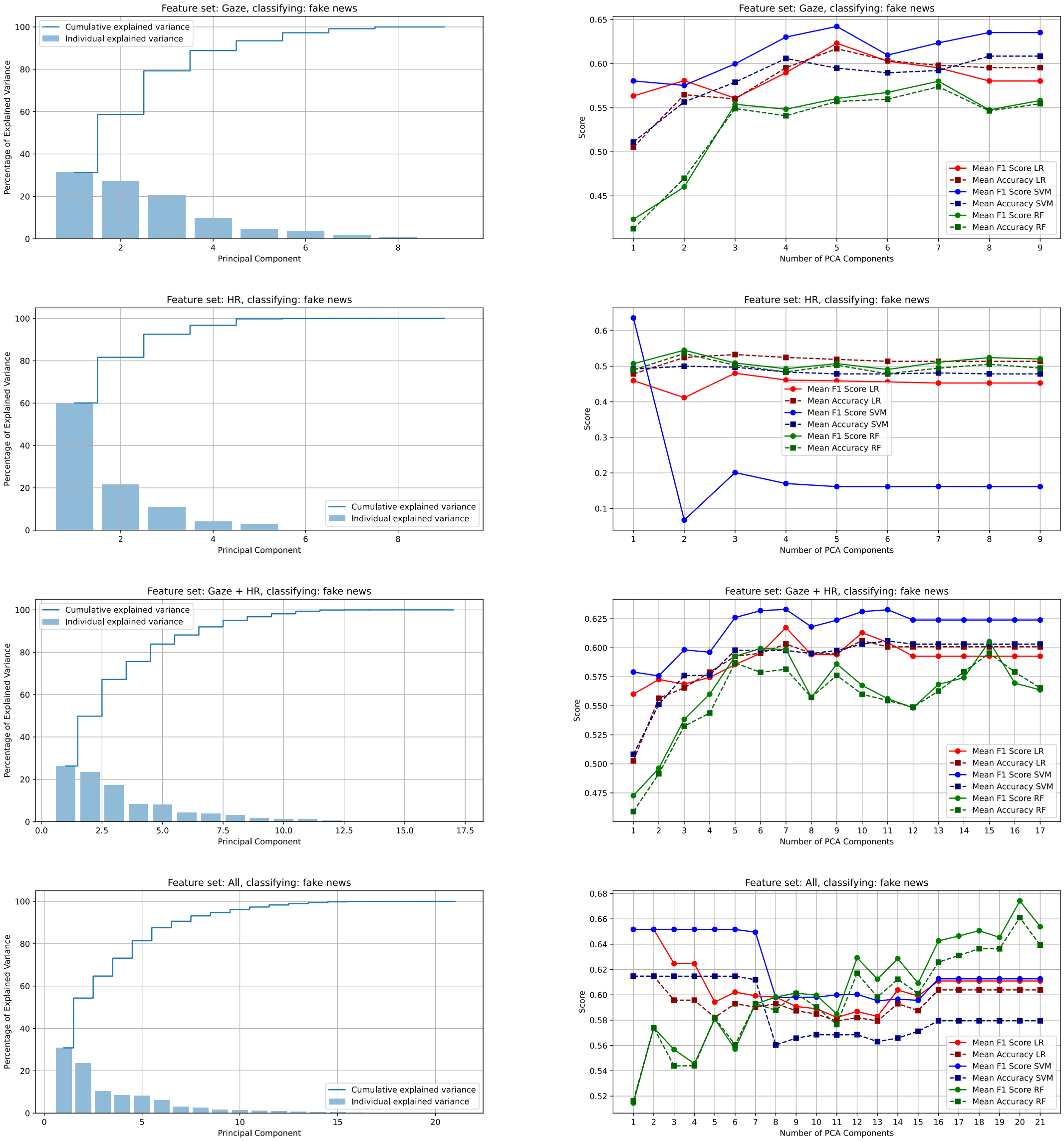
(d) Mean Pupil Diameter of Fixations

(e) HRV SDNN Ratio

(f) HRV SDSD Ratio

(g) EDA Clean Ratio

(h) HRV MeanNN Ratio



Feature set: Gaze, classifying: fake news

Feature set: HR, classifying: fake news

Feature set: Gaze + HR, classifying: fake news

Feature set: All, classifying: fake news

Figure 2: PCA analysis of predicting if news is fake or not.

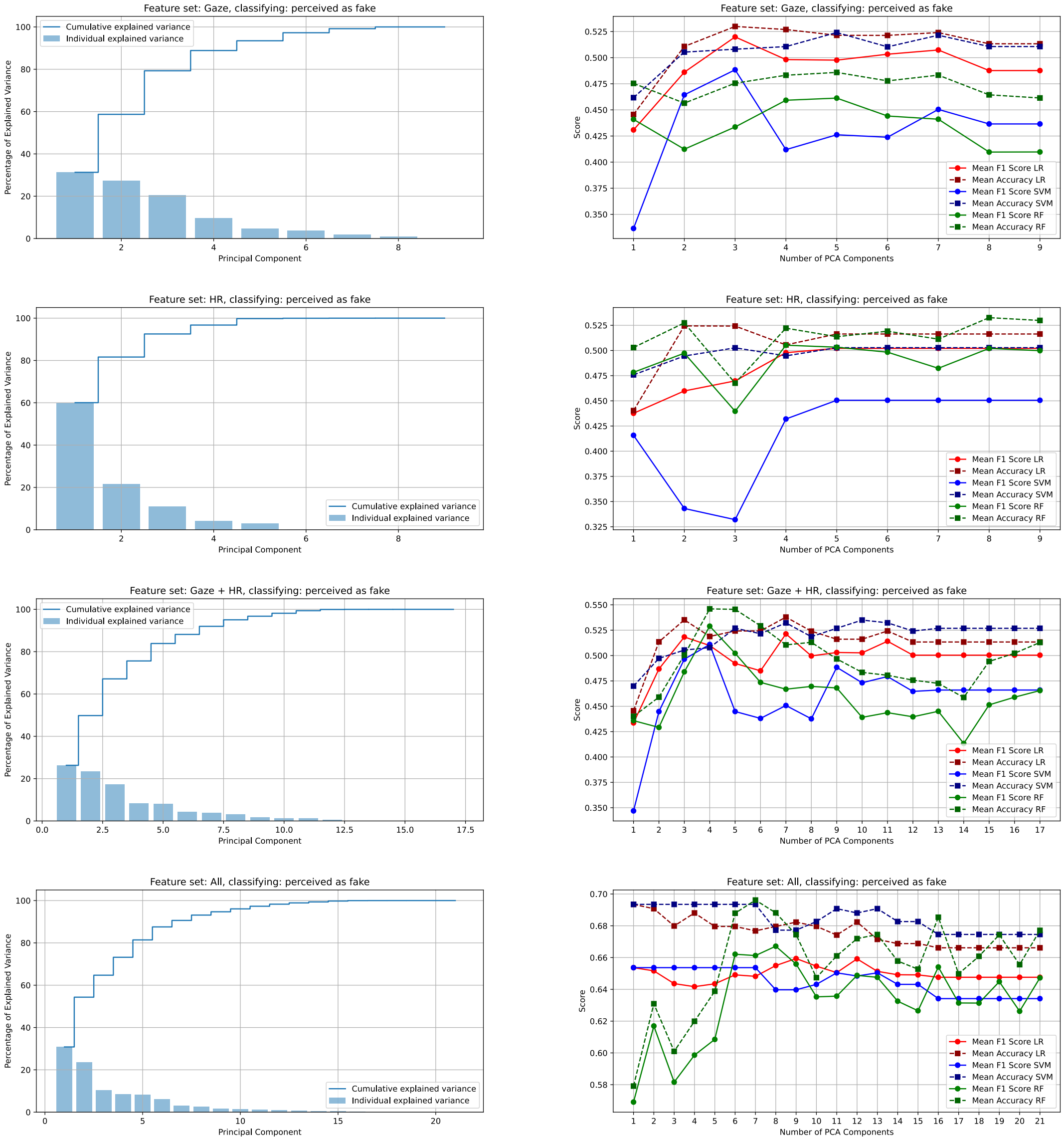


Figure 3: PCA analysis of predicting if news is perceived as fake.

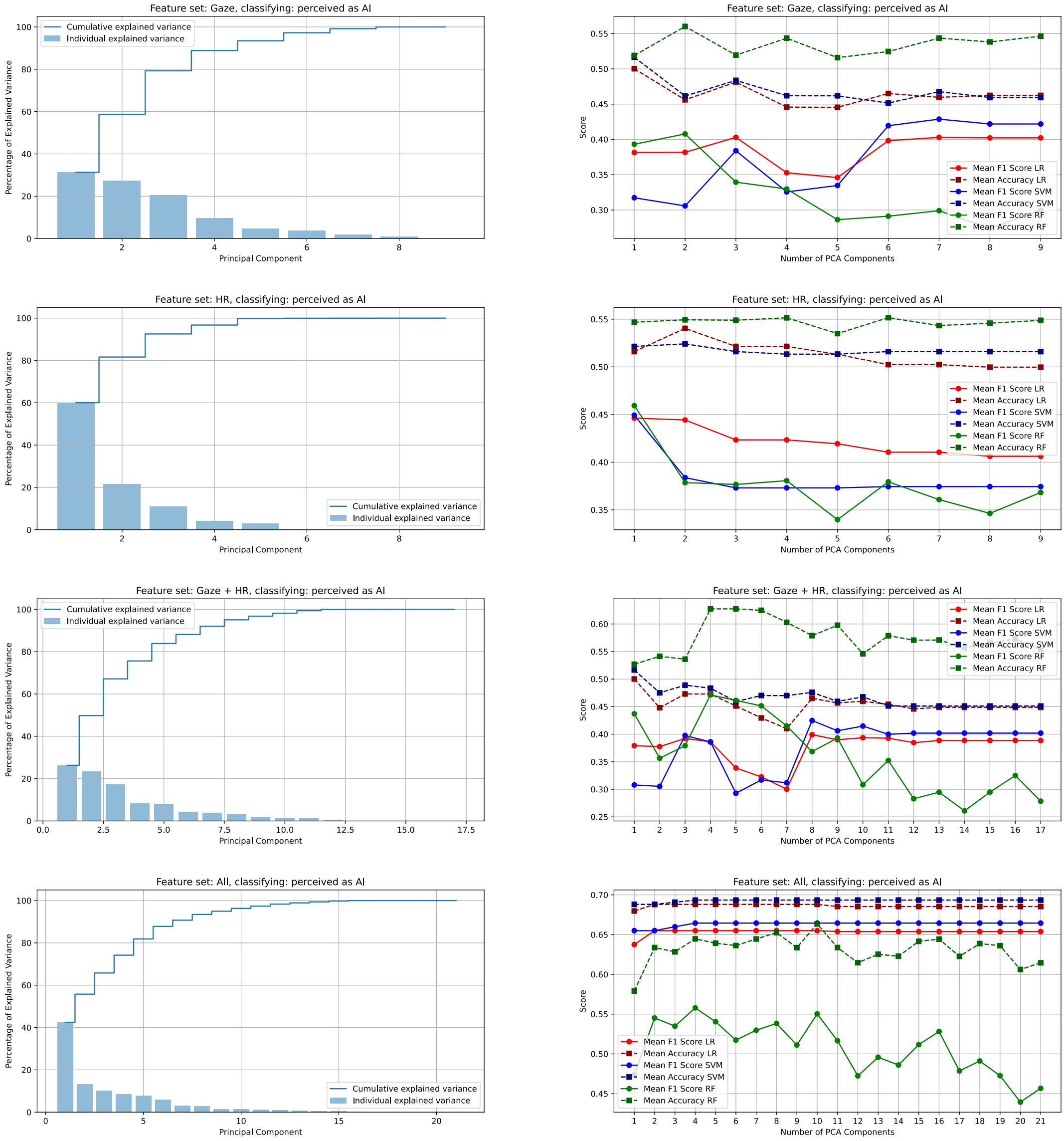


Figure 4: PCA analysis of predicting if news is perceived as AI-generated or not.

	Variable	Factor	F	df	p	
saccade	meanSaccadeDuration	TF	0.97046	1	0.32520	
		HAI	2.19772	1	0.13906	
		TF × HAI	0.95832	1	0.32824	
	regressive saccade count	TF	9.5412	1	0.0021589	**
		HAI	44.2965	1	< 0.0000	***
		TF × HAI	16.0395	1	< 0.0000	***
saccade	meanSaccadeDuration	pTF	0.089261	1	0.76528	
		pHAI	2.344527	1	0.12656	
		pTF × pHAI	0.107758	1	0.74289	
	regressive saccade count	pTF	0.013988	1	0.9059159	
		pHAI	8.687667	1	0.0034027	**
		pTF × pHAI	2.450909	1	0.1182960	
HRV	HRV SDNN	TF	0.072287	1	0.78815	
		HAI	2.046829	1	0.15316	
		TF × HAI	0.725696	1	0.39470	
	HRV SDSD	TF	0.24903	1	0.61799	
		HAI	3.10414	1	0.07872	.
		TF × HAI	1.46392	1	0.22689	
HRV	HRV SDNN	pTF	23.44208	1	< 0.0000	***
		pHAI	10.46781	1	0.0012966	**
		pTF × pHAI	0.86976	1	0.3514801	
	HRV SDSD	pTF	31.5760	1	< 0.0000	***
		pHAI	10.9423	1	0.0010088	**
		pTF × pHAI	1.5982	1	0.2067525	

Table 2: ANOVA on additional Gaze and HR features.

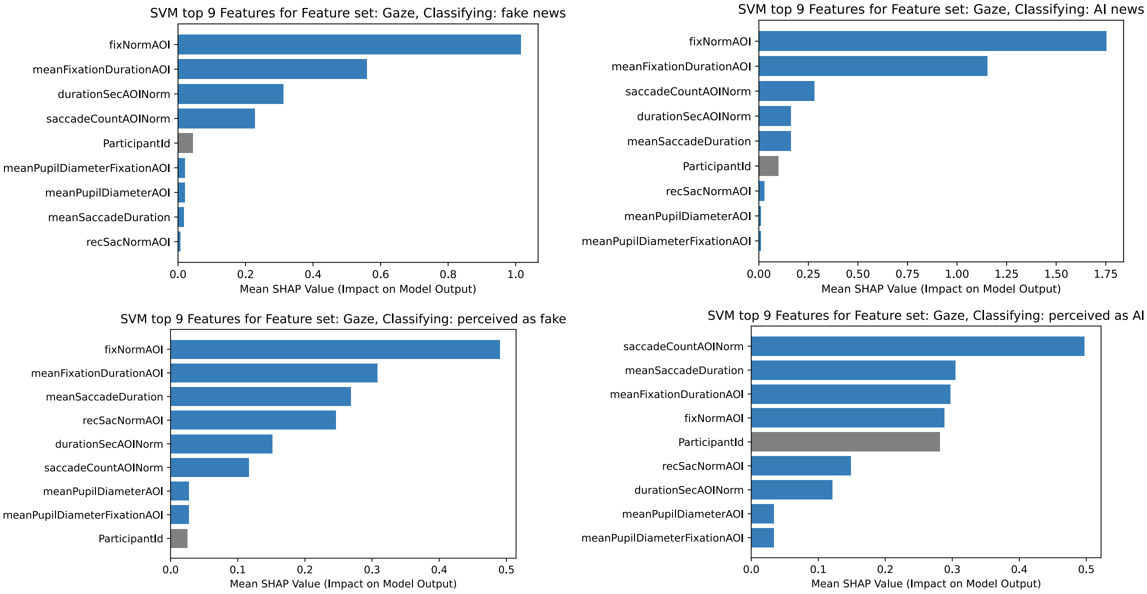


Figure 5: SHAP Values - Impact on SVM Model Output, with gaze feature set enabled

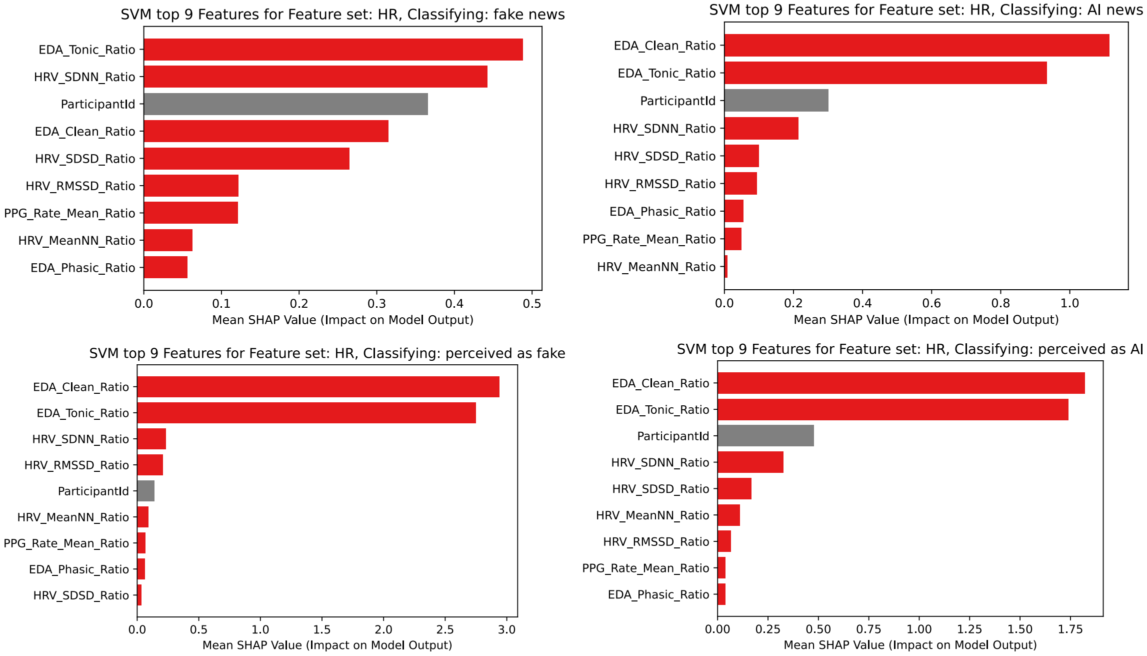


Figure 6: SHAP Values - Impact on SVM Model Output, with heart rate feature set enabled

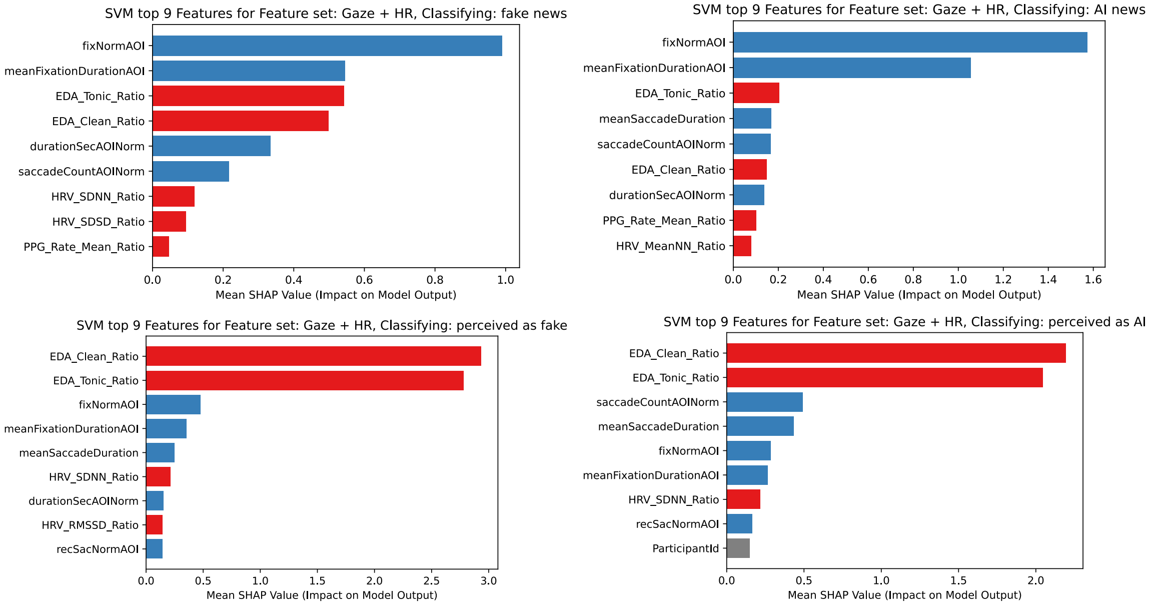


Figure 7: SHAP Values - Impact on SVM Model Output, with gaze and heart rate feature sets enabled

BIBLIOGRAPHY

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. DOI: 10.1126/science.aap9559. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aap9559.

[2] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran, "Falling for fake news: Investigating the consumption of news via social media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–10, ISBN: 9781450356206. DOI: 10.1145/3173574.3173950.

[3] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[4] J. Roozenbeek, C. R. Schneider, S. Dryhurst, *et al.*, "Susceptibility to misinformation about covid-19 around the world," *Royal Society Open Science*, vol. 7, no. 10, p. 201 199, 2020. DOI: 10.1098/rsos.201199.

[5] S. B. Naeem, R. Bhatti, and A. Khan, "An exploration of how fake news is taking over social media and putting public health at risk," *Health Information & Libraries Journal*, vol. 38, no. 2, pp. 143–149, 2021. DOI: 10.1111/hir.12320.

[6] Ö. Sümer, E. Bozkir, T. Kübler, S. Grüner, S. Utz, and E. Kasneci, "Fakenewsperception: An eye movement dataset on the perceived believability of news stories," *Data in Brief*, vol. 35, p. 106 909, 2021, ISSN: 2352-3409. DOI: 10.1016/j.dib.2021.106909.

[7] Y. Abdrabou, E. Karypidou, F. Alt, and M. Hassib, "Investigating User Behaviour Towards Fake News on Social Media Using Eye Tracking and Mouse Movements," in *Proceedings of the Usable Security Mini Conference 2023*, ser. USEC'23, abdrabou2023usec, San Diego, CA, USA: Internet Society, 2023. [Online]. Available: https://www.unibw.de/usable-security-and-privacy/publikationen/pdf/abdrabou2023usec.pdf.

[8] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, vol. 188, pp. 39–50, 2019, The Cognitive Science of Political Thought, ISSN: 0010-0277. DOI: 10.1016/j.cognition.2018.06.011.

[9] C. Martel, G. Pennycook, and D. G. Rand, "Reliance on emotion promotes belief in fake news," *Cognitive Research: Principles and Implications*, vol. 5, no. 1, p. 47, Oct. 2020, ISSN: 2365-7464. DOI: 10.1186/s41235-020-00252-3.

[10] P. M. Desmet, M. H. Vastenburg, and N. Romero, “Mood measurement with pick-a-mood: Review of current methods and design of a pictorial self-report scale,” *Journal of Design Research*, vol. 14, no. 3, pp. 241–279, 2016. DOI: 10.1504/JDR.2016.079751.

[11] M. M. Bradley and P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994, ISSN: 0005-7916. DOI: 10.1016/0005-7916(94)90063-9.

[12] J. W. Yee Chung, H. C. Fuk So, M. M. Tak Choi, V. C. Man Yan, and T. K. Shing Wong, “Artificial intelligence in education: Using heart rate variability (hrv) as a biomarker to assess emotions objectively,” *Computers and Education: Artificial Intelligence*, vol. 2, p. 100 011, 2021, ISSN: 2666-920X. DOI: 10.1016/j.caeai.2021.100011.

[13] G. Valenza, M. Nardelli, A. Lanatà, *et al.*, “Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1625–1635, 2014. DOI: 10.1109/JBHI.2013.2290382.

[14] K.-H. Choi, J. Kim, O. S. Kwon, M. J. Kim, Y. H. Ryu, and J.-E. Park, “Is heart rate variability (hrv) an adequate tool for evaluating human emotions? – a focus on the use of the international affective picture system (iaps),” *Psychiatry Research*, vol. 251, pp. 192–196, 2017, ISSN: 0165-1781. DOI: https://doi.org/10.1016/j.psychres.2017.02.025.

[15] F. García-Peñalvo and A. Vázquez-Ingelmo, “What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai,” 2023. DOI: 10.9781/ijimai.2023.07.006. [Online]. Available: https://reunir.unir.net/handle/123456789/15134.

[16] J. C. Fiona Fui-Hoon Nah Ruilin Zheng *et al.*, “Generative ai and chatgpt: Applications, challenges, and ai-human collaboration,” *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, 2023. DOI: 10.1080/15228053.2023.2233814. [Online]. Available: https://doi.org/10.1080/15228053.2023.2233814.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[19] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[20] OpenAI, “Gpt-4 technical report,” *View in Article*, vol. 2, p. 13, 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf.

[21] S. Kreps, R. M. McCain, and M. Brundage, “All the news that’s fit to fabricate: Ai-generated text as a tool of media misinformation,” *Journal of Experimental Political Science*, vol. 9, no. 1, pp. 104–117, 2022. DOI: 10.1017/XPS.2020.37.

[22] G. Spitale, N. Biller-Andorno, and F. Germani, *Ai model gpt-3 (dis)informs us better than humans*, 2023. DOI: 10.48550/arXiv.2301.11924. arXiv: 2301.11924 [cs.CY].

[23] C. Longoni, A. Fradkin, L. Cian, and G. Pennycook, “News from generative artificial intelligence is believed less,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 97–106, ISBN: 9781450393522. DOI: 10.1145/3531146.3533077.

[24] A. Graefe, M. Haim, B. Haarmann, and H.-B. Brosius, “Readers’ perception of computer-generated news: Credibility, expertise, and readability,” *Journalism*, vol. 19, no. 5, pp. 595–610, 2018. DOI: 10.1177/1464884916641269. eprint: https://doi.org/10.1177/1464884916641269. [Online]. Available: https://doi.org/10.1177/1464884916641269.

[25] Y. Yin, N. Jia, and C. J. Wakslak, “Ai can help people feel heard, but an ai label diminishes this impact,” *Proceedings of the National Academy of Sciences*, vol. 121, no. 14, e2319112121, 2024. DOI: 10.1073/pnas.2319112121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2319112121. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2319112121.

[26] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, “Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–20. DOI: 10.1145/3544548.3581318.

[27] L. Larsson, M. Nyström, R. Andersson, and M. Stridh, “Detection of fixations and smooth pursuit movements in high-speed eye-tracking data,” *Biomedical Signal Processing and Control*, vol. 18, pp. 145–152, 2015, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2014.12.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809414002031.

[28] D. D. Salvucci and J. H. Goldberg, “Identifying fixations and saccades in eye-tracking protocols,” in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ser. ETRA ’00, Palm Beach Gardens, Florida, USA: Association for Computing Machinery, 2000, pp. 71–78, ISBN: 1581132808. DOI: 10.1145/355017.355028.

[29] K. Rayner, “The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search,” *Quarterly Journal of Experimental Psychology*, vol. 62, no. 8, pp. 1457–1506, 2009, PMID: 19449261. DOI: 10.1080/17470210902816461. eprint: https://doi.org/10.1080/17470210902816461. [Online]. Available: https://doi.org/10.1080/17470210902816461.

[30] E. Matin, “Saccadic suppression: A review and an analysis.,” *Psychological bulletin*, vol. 81, no. 12, pp. 899–917, 1974. DOI: 10.1037/h0037368.

[31] K. Rayner and A. Pollatsek, “Eye movements and scene perception.,” *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 46, no. 3, pp. 342–376, 1992. DOI: 10.1037/h0084328.

[32] K. Rayner, “Eye movements in reading and information processing.,” *Psychological bulletin*, vol. 85, no. 3, pp. 618–660, 1978. DOI: 10.1037/0033-2909.85.3.618.

[33] K. Rayner, “Eye movements in reading and information processing: 20 years of research.,” *Psychological bulletin*, vol. 124, no. 3, pp. 372–422, 1998. DOI: 10.1037/0033-2909.124.3.372.

[34] J. Arizpe, D. J. Kravitz, G. Yovel, and C. I. Baker, “Start position strongly influences fixation patterns during face processing: Difficulties with eye movements as a measure of information use,” *PLOS ONE*, vol. 7, no. 2, pp. 1–17, Feb. 2012. DOI: 10.1371/journal.pone.0031106. [Online]. Available: https://doi.org/10.1371/journal.pone.0031106.

[35] Holmqvist, Kenneth and Nyström, Marcus and Andersson, Richard and Dewhurst, Richard and Halszka, Jarodzka and van de Weijer, Joost, *Eye Tracking : A Comprehensive Guide to Methods and Measures*, eng. Oxford University Press, 2011, ISBN: 9780199697083. [Online]. Available: %7Bhttp://ukcatalogue.oup.com/product/9780199697083.do%7D.

[36] K. Holmqvist, M. Nyström, and F. Mulvey, “Eye tracker data quality: What it is and how to measure it,” in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA ’12, Santa Barbara, California: Association for Computing Machinery, 2012, pp. 45–52, ISBN: 9781450312219. DOI: 10.1145/2168556.2168563. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/2168556.2168563.

[37] E. Bozkir, G. Kasneci, S. Utz, and E. Kasneci, “Regressive saccadic eye movements on fake news,” in *2022 Symposium on Eye Tracking Research and Applications*, ser. ETRA ’22, Seattle, WA, USA: Association for Computing Machinery, 2022, ISBN: 9781450392525. DOI: 10.1145/3517031.3529619.

[38] K. Shu, S. Wang, and H. Liu, “Understanding user profiles on social media for fake news detection,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 430–435. DOI: 10.1109/MIPR.2018.00092.

[39] T. Pham, Z. J. Lau, S. H. A. Chen, and D. Makowski, “Heart rate variability in psychology: A review of hrv indices and an analysis tutorial,” *Sensors*, vol. 21, no. 12, 2021, ISSN: 1424-8220. DOI: 10.3390/s21123998. [Online]. Available: https://www.mdpi.com/1424-8220/21/12/3998.

[40] G. G. BERNTSON, J. THOMAS BIGGER JR., D. L. ECKBERG, *et al.*, “Heart rate variability: Origins, methods, and interpretive caveats,” *Psychophysiology*, vol. 34, no. 6, pp. 623–648, 1997. DOI: https://doi.org/10.1111/j.1469-8986.1997.tb02140.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1997.tb02140.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1997.tb02140.x.

[41] L. Kirkwood and R. Minas, “Approaching fake news at the expense of truth: A psychophysiological study of news on social media,” 2020. [Online]. Available: http://hdl.handle.net/10125/64486.

[42] S. Pengnate, “Shocking secret you won’t believe! emotional arousal in clickbait headlines: An eye-tracking analysis,” *Online Information Review*, vol. 43, no. 7, pp. 1136–1150, 2019. DOI: 10.1108/OIR-05-2018-0172.

[43] W. Y. Wang, ““liar, liar pants on fire”: A new benchmark dataset for fake news detection,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. [Online]. Available: https://aclanthology.org/P17-2067.

[44] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media,” *Big Data*, vol. 8, no. 3, pp. 171–188, 2020, PMID: 32491943. DOI: 10.1089/big.2020.0062.

[45] “A very brief measure of the big-five personality domains,” *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, 2003, ISSN: 0092-6566. DOI: 10.1016/S0092-6566(03)00046-1.

[46] J. A. Saenz, S. R. Kalathur Gopal, and D. Shukla, *Covid-19 fake news infodemic research dataset (covid19-fnir dataset)*, 2021. DOI: 10.21227/b5bt-5244. [Online]. Available: https://dx.doi.org/10.21227/b5bt-5244.

[47] N. Dias, G. Pennycook, and D. G. Rand, “Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media,” 2020. DOI: 10.37016/mr-2020-001.

[48] A. B. Ciccone, J. A. Siedlik, J. M. Wecht, J. A. Deckert, N. D. Nguyen, and J. P. Weir, “Reminder: Rmssd and sd1 are identical heart rate variability metrics,” *Muscle & Nerve*, vol. 56, no. 4, pp. 674–678, 2017. DOI: https://doi.org/10.1002/mus.25573. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mus.25573. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mus.25573.

[49] E. Babaei, B. Tag, T. Dingler, and E. Velloso, “A critique of electrodermal activity practices at chi,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21, , Yokohama, Japan, Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445370. [Online]. Available: https://doi.org/10.1145/3411764.3445370.

[50] H. F. Posada-Quintero and K. H. Chon, “Innovations in electrodermal activity data collection and signal processing: A systematic review,” en, *Sensors (Basel)*, vol. 20, no. 2, Jan. 2020.

[51] Tobii Technology, *Tobii pro fusion: Reach further with your research*, 2024. [Online]. Available: https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion.

[52] Empatica Inc, *E4 sensors*, 2024. [Online]. Available: https://www.empatica.com/research/e4/.

[53] A. Savitzky and M. J. E. Golay, “Smoothing and differentiation of data by simplified least squares procedures.,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964. DOI: 10.1021/ac60214a047. eprint: https://doi.org/10.1021/ac60214a047. [Online]. Available: https://doi.org/10.1021/ac60214a047.

[54] R. W. Schafer, “What is a savitzky-golay filter? [lecture notes],” *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111–117, 2011. DOI: 10.1109/MSP.2011.941097.

[55] N. Nachar *et al.*, “The mann-whitney u: A test for assessing whether two independent samples come from the same distribution,” 1, vol. 4, 2008, pp. 13–20. [Online]. Available: https://api.semanticscholar.org/CorpusID:59357756.

[56] H. Gil de Zúñiga, B. Weeks, and A. Ardèvol-Abreu, “Effects of the News-Finds-Me Perception in Communication: Social Media Use Implications for News Seeking and Learning About Politics,” *Journal of Computer-Mediated Communication*, vol. 22, no. 3, pp. 105–123, Apr. 2017, ISSN: 1083-6101. DOI: 10.1111/jcc4.12185. eprint: https://academic.oup.com/jcmc/article-pdf/22/3/105/19946815/jjcmcom0105.pdf. [Online]. Available: https://doi.org/10.1111/jcc4.12185.

[57] S. A. Jessup, T. R. Schneider, G. M. Alarcon, T. J. Ryan, and A. Capiola, “The measurement of the propensity to trust automation,” in *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, J. Y. Chen and G. Fragomeni, Eds., Cham: Springer International Publishing, 2019, pp. 476–489, ISBN: 978-3-030-21565-1. DOI: 10.1007/978-3-030-21565-1_32.

[58] T. R. Schneider, S. A. Jessup, C. Stokes, S. Rivers, M. Lohani, and M. McCoy, “The influence of trust propensity on behavioral trust,” in *Poster session presented at the meeting of Association for Psychological Society, Boston*, 2017.

[59] S. Frederick, “Cognitive reflection and decision making,” *Journal of Economic Perspectives*, vol. 19, no. 4, pp. 25–42, Dec. 2005. DOI: 10.1257/089533005775196732. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/089533005775196732.

[60] M. E. Toplak, R. F. West, and K. E. Stanovich, “The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks,” *Memory & Cognition*, vol. 39, no. 7, pp. 1275–1289, Oct. 2011, ISSN: 1532-5946. DOI: 10.3758/s13421-011-0104-1. [Online]. Available: https://doi.org/10.3758/s13421-011-0104-1.

[61] O. P. John, E. M. Donahue, and R. L. Kentle, "Big five inventory," *Journal of Personality and Social Psychology*, 1991. DOI: 10.1037/t07550-000.

[62] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007, ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2006.02.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656606000195.

[63] T. Technology, *Accuracy and precision test method for remote eye trackers*, Feb. 2011.

[64] J. V. Bradley, "Complete counterbalancing of immediate sequential effects in a latin square design," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 525–528, 1958. DOI: 10.1080/01621459.1958.10501456. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501456. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501456.

[65] D. Makowski, T. Pham, Z. J. Lau, *et al.*, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, Feb. 2021. DOI: 10.3758/s13428-020-01516-y. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y.

[66] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, , Vancouver, BC, Canada, Association for Computing Machinery, 2011, pp. 143–146, ISBN: 9781450302289. DOI: 10.1145/1978942.1978963. [Online]. Available: https://doi.org/10.1145/1978942.1978963.

[67] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993, ISSN: 0098-3004. DOI: https://doi.org/10.1016/0098-3004(93)90090-R. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009830049390090R.

[68] S. M. Holland, "Principal components analysis (pca)," *Department of Geology, University of Georgia, Athens, GA*, vol. 30602, p. 2501, 2008. [Online]. Available: http://stratigrafia.org/8370/handouts/pcaTutorial.pdf.

[69] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014. DOI: https://doi.org/10.48550/arXiv.1404.1100.

[70] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[71] S. Hutton, “Cognitive control of saccadic eye movements,” *Brain and Cognition*, vol. 68, no. 3, pp. 327–340, 2008, A Hundred Years of Eye Movement Research in Psychiatry, ISSN: 0278-2626. DOI: https://doi.org/10.1016/j.bandc.2008.08.021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278262608002662.

[72] J. Zagermann, U. Pfeil, and H. Reiterer, “Studying eye movements as a basis for measuring cognitive load,” in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA ’18, Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–6, ISBN: 9781450356213. DOI: 10.1145/3170427.3188628. [Online]. Available: https://doi.org/10.1145/3170427.3188628.

[73] D. Long and B. Magerko, “What is ai literacy? competencies and design considerations,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20, Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–16, ISBN: 9781450367080. DOI: 10.1145/3313831.3376727. [Online]. Available: https://doi.org/10.1145/3313831.3376727.

[74] M. Veale and F. Z. Borgesius, “Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach,” *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021. DOI: doi:10.9785/cri-2021-220402. [Online]. Available: https://doi.org/10.9785/cri-2021-220402.

[75] K. Roy and D. Chanda, “A robust webcam-based eye gaze estimation system for human-computer interaction,” in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2022, pp. 146–151. DOI: 10.1109/ICISET54810.2022.9775896.

[76] M. S. Mounica, M. Manvita, C. Jyotsna, and J. Amudha, “Low cost eye gaze tracker using web camera,” in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 79–85. DOI: 10.1109/ICCMC.2019.8819645.