

# WARSAW UNIVERSITY OF TECHNOLOGY

Discipline of Science Engineering and Technology  
Field of Science Information and Communication Technology

## Ph.D. Thesis

Karolina Gabor-Siatkowska, M.Sc.

**Improving Therapeutic Spoken Dialogue Systems  
with Eye Tracking**  
(Ulepszanie terapeutycznych głosowych systemów dialogowych  
przy użyciu okulografii)

**Supervisor:**  
Prof. Artur Janicki, Ph.D., D.Sc.

Warsaw 2025

## Acknowledgements

I am incredibly thankful to my supervisor, Prof. Artur Janicki, for his guidance throughout my PhD journey and for every constructive criticism. His support and solution-oriented perspective have shaped not only my work but also the way I approach new challenges. I am also deeply grateful to him for sharing his great passion for science with me. Thank you for being such an inspiring mentor.

My heartfelt thanks also belong to the psychiatrist who worked on Terabot, Dr. Izabela Stefaniak, PhD, MD. I am very appreciative of the extra hours of assistance at the psychiatric hospital for me (thanks to which I could collect more data). Her openness and enthusiasm for technological solutions in psychiatry are invaluable. I truly believe that not only does the scientific world need more medical doctors like her.

While writing this dissertation, I have to mention the people closest to me. These people are my husband and my parents. They were with me during every good and bad moment while I was completing my doctorate. Thanks to their constant support and unwavering belief in me, even when I lacked it myself, it would not have been possible. I cannot put into words how grateful I am for their support.

I would also like to thank the “Teatr PW” theater Director, Grzegorz Sierzputowski, for his wholehearted enthusiasm for participation in the spoken dialogue system tests, and to the “Teatr PW” theater Students for executing them with such passion.

I appreciate the opportunity to have contributed to a society-meaningful IT project that aimed to support patients with severe mental health issues. The experience I gained in my PhD journey made me realise that science is not just about analyzing data, designing systems, or solving technical problems. It is also a great way to share knowledge, enjoy discovering new ideas, and put them into practice.

I believe this is a very valuable lesson that will hopefully shape my future.

## Streszczenie

Tematem niniejszej rozprawy jest wykorzystanie technologii śledzenia ruchu gałek ocznych (okulografii) w celu udoskonalenia głosowego systemu dialogowego do celów terapeutycznych. System ten został stworzony na potrzeby pacjentów z rozpoznaniem schizofrenii. Podczas rozmów pacjentów z Terabotem, wystąpił szereg problemów, głównie związanych z płynnością dialogu. W celu znalezienia odpowiednich rozwiązań, wykorzystano dane okulograficzne.

Obecnie okulografia jest szeroko wykorzystywana zarówno w badaniach naukowych jak i zastosowaniach komercyjnych obejmujących różnorodne dziedziny, np. medycynę, edukację, psychologię czy marketing. Przykładem zastosowania w dziedzinie "komunikacji człowiek – komputer" są tzw. agenty konwersacyjne lub systemy dialogowe, którym poświęca się ostatnio wiele uwagi. Jednym z zastosowań jest np. wykorzystanie zachowań niewerbalnych do nawiązania kontaktu z użytkownikiem. Kluczową koncepcją w projektowaniu tych systemów jest wykorzystanie zachowań niewerbalnych do zbierania informacji o użytkownikach, zamiast polegania wyłącznie na mowie. Aby dialog między agentem a człowiekiem stał się bardziej realistyczny, agenci mogą być wyposażeni w zdolności komunikacyjne i ekspresyjne podobne do tych obserwowanych w interakcji z ludźmi, w czasie której mamy do czynienia z ich mową, gestami mimiką twarzy czy wzrokiem). Dzięki temu dialog może przebiegać bardziej naturalnie, a systemy mogą zapewnić bardziej realistyczną rozmowę z użytkownikami. Dane ze śledzenia ruchu gałek ocznych mogą służyć jako dodatkowe źródło informacji, szczególnie podczas rozmów z agentami dialogowymi w aspekcie zdrowia psychicznego.

Terapeutyczny system dialogowy Terabot został zaprojektowany przez zespół badawczy z Politechniki Warszawskiej (PW), Wydziału Elektroniki i Technik Informacyjnych. Jego utworzenie miało na celu wzbogacenie terapii pacjentów z zaburzeniami takimi jak np. depresja, zaburzenia lękowe itp. W celu zweryfikowania skuteczności systemu przeprowadzono pilotażowe randomizowane badanie kliniczne w Instytucie Psychiatrii i Neurologii w Warszawie. 38 pacjentów miało możliwość przeprowadzania rozmów terapeutycznych z systemem dialogowym Terabot (oprócz bieżącego leczenia, w tym farmakologicznego). Podczas rozmowy z Terabotem pacjent mógł wybrać jedną z trzech emocji: lęk, strach lub wstyd. Pod koniec rozmowy Terabot zachęcał do wykonania ćwiczenia relaksacyjnego, które miało pomóc pacjentom w uspokojeniu się i opanowaniu emocji. Zbierane były dane okulograficzne oraz nagrania audio z dialogów pomiędzy pacjentami a Terabotem. Bezpośredni kontakt z pacjentami pozwolił na znacznie efektywniejszą optymalizację systemu dialogowego.

Na podstawie analiz zweryfikowano dotychczasowe działanie istniejącego systemu dialogowego. Przeprowadzono analizy interfejsu graficznego Terabota, w szczególności obszarów zainteresowania pacjentów (ang. area of interest, AOI). Analizowano gdzie pacjenci patrzyli najdłużej i najczęściej podczas gdy kończyli swoją wypowiedź. Wyniki analiz stały się kluczowym źródłem informacji, które posłużyły do zaprojektowania kolejnej wersji systemu. Nowy system jest systemem multimodalnym, wspomaganym przez sygnał okulograficzny. Jest on

w stanie reagować na zachowanie pacjenta (jego spojrzenie) podczas rozmowy. Ponadto system ten opiera się na dużych modelach językowych (ang. Large Language Models, LLM), w szczególności na jednym z modeli GPT firmy OpenAI. Takie podejście sprawia, że nowy system jest bardziej zaawansowany nie tylko naukowo, ale i technologicznie. Dzięki zastosowaniu okulografu (w przeciwieństwie do systemu dialogowego bez dodatkowych narzędzi) można było dostosować konwersacje do zachowania pacjenta. Pozwalało to na bardziej naturalną interakcję między pacjentem a systemem dialogowym. Dane okulograficzne posłużyły do rozwiązania problemów, które powodowały przerwy w dialogu między systemem a pacjentem. Ponadto technologię śledzenia ruchu gałek ocznych wykorzystano również do automatycznej oceny zaangażowania pacjentów w dialog. Proponowane rozwiązania mogą mieć pozytywny wpływ na interakcję pomiędzy Terabotem a pacjentem. Może się to również przyczynić się do zwiększenia dobrego samopoczucia pacjentów i komfortu podczas sesji terapeutycznych.

*Słowa kluczowe:* okulografia, system dialogowy, interakcja człowiek – komputer, systemy multimodalne, psychiatria, komputerowe wspomaganie medycyny, terapia poznawczo - behawioralna, fiksacje, duże modele językowe, GPT, automatyczna ocena zaangażowania, brak kalibracji

# Abstract

The subject of this dissertation focuses on using eye tracking to improve a spoken dialogue system for therapeutic purposes. It was used during a pilot clinical trial, during which it was tested on patients diagnosed with schizophrenia. Problems in dialogue flow were identified. Solutions were found by using eye-tracking data.

Today, eye tracking is widely used in science and commercial applications in various fields such as medicine, education, psychology, and marketing. Embodied conversational agents (ECAs), which use eye-tracking data, are an example of an application in the field of human-computer interaction that has received much attention in recent years. The key concept in designing these systems is to use non-verbal behavior to collect information about users, rather than relying purely on speech. To make the dialogue between the agent and the human more realistic, these agents can be equipped with communication and expressive capabilities similar to those observed in human interaction (e.g., speech, gestures, facial expressions, and gaze). This can make the dialogue flow more naturally, and therefore, the systems can provide a more realistic conversation with users. Especially when it comes to the digital mental health domain, eye-tracking data might be helpful as an additional source of information during conversations with patients.

The therapeutic spoken dialogue system, Terabot, was designed by a research team at the Warsaw University of Technology (WUT), Faculty of Electronics and Information Technology. It was designed with the aim of serving as an addition to the treatment of mental disorders (e.g., depression, anxiety disorders). In order to verify the effectiveness of the system, it was implemented at the Institute of Psychiatry and Neurology (IPIN) in Warsaw. A pilot clinical trial was conducted, in which patients diagnosed with schizophrenia could talk with Terabot as part of their ongoing treatment. Eye-tracking and audio data from dialogues between patients and Terabot were collected during these sessions. The specifics of patients' behavior during conversations with the dialogue system were considered, and based on them, the performance of the existing Terabot system was verified. Analyses of Terabot's interface were conducted to identify the areas of interest to patients, at which they mostly gazed while they concluded their speaking. These results provided valuable information for designing a new, multimodal version of the system: an LLM-based, gaze-enhanced spoken dialogue system. It is capable of responding to the patient's behavior (indicated by gaze) during conversation. In addition, the new system is based on large language models (LLMs), on one of the GPT models by OpenAI. This approach made the new system scientifically and technologically more advanced. By using the eye tracker (as opposed to the dialogue system without additional tools), the conversations have adapted to the patient's gaze behavior. This approach allows the conversation between the patient and the dialogue system to be more natural and enables the system to handle situations in which speech-only would set the whole dialogue on pause. In addition, eye tracking was used to assess patient engagement in the dialogue.

*Keywords:* eye tracking, dialogue agent, human-computer interaction, multimodal systems, psychiatry, computer-aided therapy, cognitive-behavioral therapy, fixations, gaze, large language models, GPT, automatic patients' engagement, calibration-free

# List of contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	General background .....	1
1.2	Scope of this thesis.....	4
1.3	My publications and public presentations .....	6
1.4	Participation in research projects.....	8
1.5	Other achievements .....	10
<b>2</b>	<b>Literature review</b>	<b>11</b>
2.1	Mental health care as a global problem .....	11
2.2	Dialogue (conversational) agents .....	13
2.2.1	Main components of spoken dialogue systems .....	14
2.2.2	Dialogue systems applied in the domain of digital mental health .....	16
2.2.3	LLM-based solutions for psychological well-being and mental health ...	21
2.3	Eye tracking .....	23
2.3.1	Technical parameters of eye trackers.....	26
2.3.2	Eye tracking in Human–Computer–Interaction .....	30
<b>3</b>	<b>Terabot system</b>	<b>33</b>
3.1	Terabot – a RASA-based dialogue system in the Polish language for patients with psychiatric disorders .....	33
3.1.1	User interface.....	33
3.1.2	Architecture .....	34
3.2	Experiments at the Institute of Psychiatry and Neurology .....	36
3.2.1	Eye tracker Gazepoint GP3 .....	38
3.3	Problems identified during conversations .....	41
3.3.1	Issue 1: Interrupting the patients’ answers.....	42
3.3.2	Issue 2: Patients having to wait a long time for Terabot’s response .....	43
3.3.3	Issue 3: Waiting too long for the patient’s answer.....	43
3.3.4	Issue 4: Lack of information about patient behavior .....	44
3.4	Patients’ satisfaction regarding conversations with Terabot .....	45
<b>4</b>	<b>Analyzing the need for eye tracker calibration in a spoken dialogue system</b>	<b>46</b>

4.1	Problem description.....	46
4.2	Experiment methodology.....	46
4.3	Results and discussion.....	50
4.4	Conclusion.....	52
<b>5</b>	<b>Analysis of the gaze of a patient conversing with Terabot</b>	<b>53</b>
5.1	Problem description.....	53
5.2	Methodology.....	53
5.3	Utterance and eye-tracking data analysis.....	54
5.4	Results.....	57
5.5	Clinical use of the proposed solution.....	58
<b>6</b>	<b>Designing a gaze-dependent, LLM-based dialogue system for therapeutic purposes</b>	<b>60</b>
6.1	Problem description.....	60
6.2	Analysis of patients' waiting time.....	60
6.3	Using real-time eye-tracking data to improve the dialogue system.....	61
6.3.1	Using OpenAI's Assistant API for dialogue management.....	61
6.3.2	Spoken dialogue system with eye-tracking data input.....	63
6.4	Offline tests – based on patients' data.....	64
6.5	Real-time tests.....	67
6.5.1	Participants: simulated patients and “Teatr PW” theater actors.....	67
6.5.2	Experimental setup and methodology.....	68
6.6	Analysis of the new dialogue system performance.....	69
6.6.1	Eye-tracking response activation.....	70
6.6.2	Analysis of conversations.....	70
6.6.3	Survey on participants' satisfaction regarding conversations with Terabot.....	73
6.7	Conclusion.....	74
<b>7</b>	<b>Automatic assessment of patient engagement</b>	<b>75</b>
7.1	Problem description.....	75
7.2	Analysis of assistants' notes on patient engagement during dialogues.....	76
7.3	Proposed method.....	78
7.4	Results.....	79
7.5	Conclusion.....	83
<b>8</b>	<b>Summary and discussion</b>	<b>84</b>
8.1	Limitations.....	87
8.2	Further work.....	87
<b>A</b>	<b>Dialogue sample – chosen topic: anger</b>	<b>89</b>

<b>B Scenarios for actresses and actors of “Teatr PW” theater</b>	<b>91</b>
<b>List of Figures</b>	<b>94</b>
<b>List of Tables</b>	<b>97</b>

# Chapter 1

## Introduction

### 1.1 General background

Today's societies are facing a severe mental health crisis. An increasing number of people are struggling with anxiety and depression [1]. Unfortunately, there are not enough professionals available to help. This is a significant issue because it is well known that mental health directly affects one's personal life and can result in work-related disability. On top of that, the number of children and young people requiring mental healthcare has also increased in recent years. Many of them face long waiting periods before receiving proper treatment. It is especially dangerous when some of them deteriorate to a severe state of mental illness while waiting [2]. Researchers concluded that globally, the number of children and young people (aged 10–24 years) affected by depression and anxiety disorders increased over the years 1990–2021, accelerating significantly from 2014 to 2021 and culminating in 2019 [1].

Research teams around the world are working to find effective solutions. Various aspects of mental health are being investigated from different perspectives and research disciplines. Table 1 shows the various technologies applied to the digital mental health domain such as apps that use mindfulness techniques, focus on mood tracking, and provide psychoeducation; chatbots that interact with users through conversation; and social robots that analyze users' facial expressions so that their interaction is better suited to the user's emotions.

Other examples of these solutions are studies of verbal and non-verbal behavior in human–human and human–machine interactions, focusing, for example, on the detection of depression in speech signals, including those of children. There are also projects which concentrate on Parenting for Lifelong Health (PLH) programmes in low- and middle-income countries. These programmes aim to improve childcare in Southeast Asia, Eastern Europe, and sub-Saharan Africa by addressing parental mental health. Other scientific fields investigate digital mental health interventions with the aim of creating scalable, evidence-based eHealth solutions for adolescents and students. There are also many other projects already in progress [4].

In 2025, a survey was published on the different delivery methods of interventions across

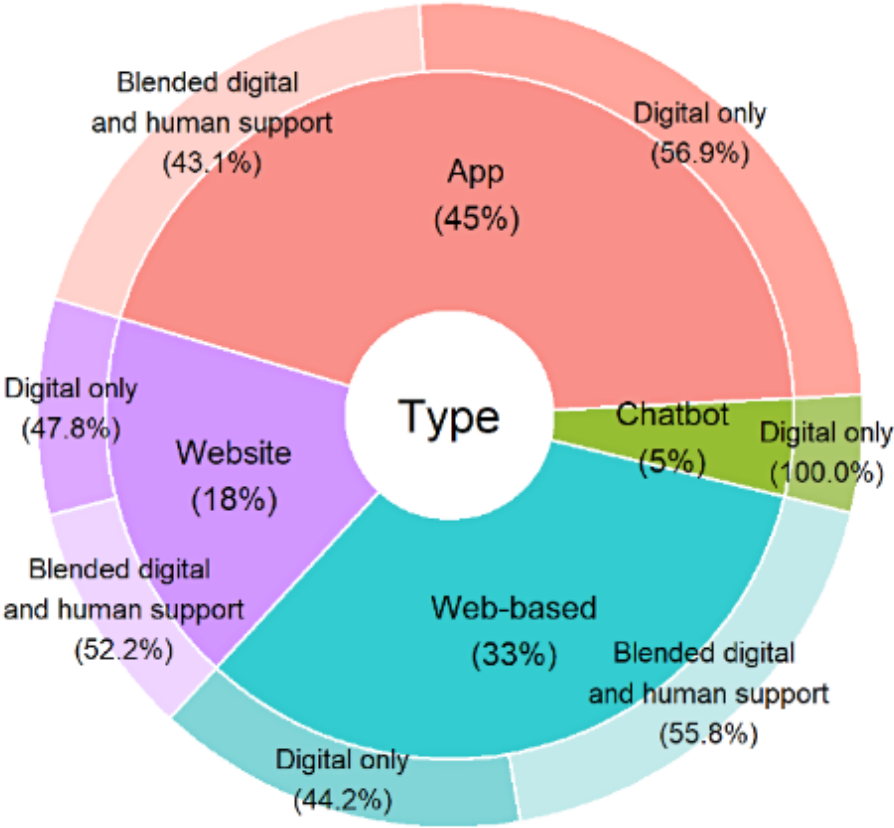
**Table 1:** Various digital technologies that can be used in mental health services, based on [3]

Digital health apps	Virtual reality	Natural language processing	Data Science	Supervised Machine Learning	Robotics and Sensing
Psychoeducation	Exposure therapy	Chatbots	Process mining	Predicting outcomes	Computer vision (e.g., Facial expression analysis)
Mood tracking	Simulation based training	Smart speakers / speech analytics	Clustering / digital phenotyping	Triaging	Psycho-physiology / Affective computing
Mindfulness	Empathy machines	Semi-automated digital counselling	Association analysis		Social robots
Digital diaries		Typed therapy	Time series analytics		Emotional Faces Memory Task
Peer support / social networking					
Patient management					

various types of digital tools (see Figure 1) [5]. The aim was to explore the range of digital mental health interventions available for young people (aged 16–25 years) with a strong focus on digital tool types, modalities, delivery formats, etc. As can be seen, most solutions focus on apps (45%), sometimes also with human support (43.1%). Another big group is represented by web-based applications (33%), where again over half of them are supported by humans. The graph indicates that chatbots are not common (only 5%). As will be further described in the dissertation, this term is often referred to as a dialogue system with which humans can interact. They are presumed to be only text-based (e.g., the Messenger application by Facebook). In contrast, there are also other technological solutions, with a combination of, for example, text-input and speech-output, or they may be entirely speech-to-speech based. The rise of artificial intelligence (AI), has also led to it being introduced into the mental health domain. In [6], the authors describe AI-based conversational agents from different countries; by 2023, 35 of these agents had been identified as connected to the digital mental health topic. Some of the main chatbots, dialogue systems, conversational agents, etc., will be described in detail in the following chapters.

The development of a proposed solution for digital mental health has many benefits, but also poses significant challenges. Many mental health apps are available, but most haven't been assessed or accredited by the relevant organizations and experts [3]; thus opinions on digital mental health solutions differ. Nevertheless, one of the objectives of digital mental health solu-

tions is not to replace the work of professionals such as psychiatrists and psychotherapists [3], instead, the objective is to develop tools that can help them, for example, by saving time. The idea is to improve their work by enabling them to focus on tasks that cannot (or should not) be replicated by any algorithm or machine, while at the same time allowing all other processes to be automated.



**Figure 1:** Diagram showing types of delivery of interventions across different types of digital tools (N = 135) [5]

## 1.2 Scope of this thesis

The thesis is stated as follows:

### **The use of eye tracking can improve a therapeutic spoken dialogue system.**

The focus is on research conducted with Terabot, a spoken dialogue system, for therapeutic purposes. It was designed by a research team of the Warsaw University of Technology (WUT), Faculty of Electronics and Information Technology. This research was funded by the Center for Priority Research Area Artificial Intelligence and Robotics of the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) program.

Terabot was designed with the aim of serving as an addition to the treatment of mental disorders (e.g. depression, anxiety disorders). The system is goal-oriented, meaning the main task was to talk with patients about one of three emotions: anxiety, fear, or shame. At the end of the conversation, a relaxation exercise was recommended to help calm the emotions. All of Terabot's answers were checked and approved by a professional (psychiatrist and psychotherapist). Terabot's dialogue sequence used elements of Cognitive Behavioral therapy (CBT) and followed psychiatric recommendations (meaning Terabot's utterances were based on CBT).

At the Institute of Psychiatry and Neurology (IPIN) in Warsaw, patients diagnosed with schizophrenia could talk with Terabot as part of their ongoing treatment. During these conversations, some unforeseen problems occurred, which paused the dialogue flow and sometimes caused unintended patient discomfort. The following problematic situations were identified when patients interacted with the dialogue system:

- Interrupting the patients' answers by Terabot,
- Letting patients wait very long for Terabot's response,
- Waiting too long for the patient's answer,
- No non-speech related information about patient participation during conversation.

Those issues are described in detail in this dissertation. While patients conversed with Terabot, audio and eye-tracking data were collected and analyzed. While investigating these different problems, the following research objectives were defined:

RO1: Investigating whether an eye tracker needs to be calibrated when used in a spoken dialogue system.

RO2: Analysis of areas of interest in the graphical interface of the dialogue system.

RO3: Enhancement of fluency of conversations with a dialogue system by using an eye tracker

RO4: Enabling automatic user engagement assessment while interacting with the dialogue system.

I used the results of the collected eye-tracking and audio data to find solutions on how to improve the therapeutic spoken dialogue system by using the eye tracker. Based on this experience, I designed a new, multimodal version of Terabot: an LLM-based dialogue system enhanced with real-time eye tracking data. This system has already been tested on healthy users.

## 1.3 My publications and public presentations

During my PhD studies, I managed to publish the following conference and journal articles:

1. R. S. Gomolka, A. Korzyska, K. Siemion, K. Gabor-Siatkowska, and W. Klonowski, “Automatic method for assessment of proliferation index in digital images of DLBCL tissue section,” *Biocybernetics and Biomedical Engineering*, vol. 39, no. 1, pp. 30–37, 2019 (IF: 2.429; 140 points MNiSW).
2. K. Gabor-Siatkowska and A. Janicki, “Czy kalibracja okulografu w interfejsie człowiek-komputer zawsze jest niezbędna?” in *Konferencja Krajowego Środowiska Tele- i Radiokomunikacyjnego*, Warsaw, 2022, *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*, ISSN 1230-3496 (20 points MNiSW).
3. K. Gabor-Siatkowska and A. Janicki, “Badanie zasadności używania kalibracji w pomiarach okulograficznych w celu zastosowania w interfejsie człowiek-komputer,” *Rozwiązania technologiczne XXI wieku – skutki i perspektywy rozwoju*. Tom 3, 2023, Wydawnictwo Naukowe TYGIEL Sp. z o. o., 160 s., ISBN 978-83-67104-75-3 (20 points MNiSW).
4. M. Kozłowski, K. Gabor-Siatkowska, I. Stefaniak, M. Sowański, and A. Janicki, “Enhanced emotion and sentiment recognition for empathetic dialogue system using big data and deep learning methods,” in the *International Conference on Computational Science (ICCS 2023)*, Prague, Czech Republic, 3-5 July 2023, Springer, 2023, pp. 465–480 (140 points MNiSW).
5. K. Gabor-Siatkowska, M. Sowański, M. Pudo, R. Rzatkiwicz, I. Stefaniak, M. Kozłowski, and A. Janicki, “Therapeutic spoken dialogue system in clinical settings: Initial experiments,” in *30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, in Ohrid, North Macedonia. IEEE, 2023, pp. 1–5 (20 points MNiSW).
6. K. Gabor-Siatkowska, M. Sowański, R. Rzatkiwicz, I. Stefaniak, M. Kozłowski, and A. Janicki, “AI to train AI: Using ChatGPT to improve the accuracy of a therapeutic dialogue system,” *Electronics*, vol. 12, no. 22, p. 4694, 2023 (IF: 2.6, 140 points MNiSW).
7. K. Gabor-Siatkowska, I. Stefaniak, and A. Janicki, “Gaze-dependent response activation in dialogue agent for cognitive-behavioral therapy,” in the *28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)*, 2024, 11-09-2024 - 13-09-2024, Sevilla, Spain, *Procedia Computer Science*, 2024, vol. 246, pp. 2322–2331, (70 points MNiSW).
8. K. Gabor-Siatkowska, I. Stefaniak, and A. Janicki, “Eye tracking data cleansing for dialogue agent,” in “*Sztucznej Inteligencji - aplikacje i implikacje*” Wrocław University of Applied Informatics, *Biuletyn Naukowy Wrocławskiej Wyższej Szkoły Informatyki Stosowanej Informatyka*, vol. 10, pp. 1–14, 2024, ISSN 2082-9892 (5 points MNiSW).

9. K. Gabor-Siatkowska, I. Stefaniak, and A. Janicki, “A multimodal approach for improving a dialogue agent for therapeutic sessions in psychiatry” in *Transforming Media Accessibility in Europe*. Springer, 2024, pp. 397–414 (20 points MNiSW).
10. K. Gabor-Siatkowska, I. Stefaniak, and A. Janicki, “Beyond Words: Gaze-Enhanced LLM-based Dialogue System for Therapeutic Purposes”, in the 32nd International Conference on Systems, Signals and Image Processing (IWSSIP 2025), Skopje, North Macedonia, IEEE, 2025, pp. 1–5, (20 points MNiSW).

Apart from presenting my work at scientific conferences, I also had the chance to present my achievements at international seminars and scientific events:

1. Presentation of scientific results at an interdisciplinary, international seminar on the topic “Media Accessibility in the Age of Artificial Intelligence” of European Cooperation in Science and Technology COST action (CA19142) in Limassol, Cyprus (04.06.2024).
2. Presentation of the published chapter in the international book at the closing conference of the COST (European Cooperation in Science and Technology) international action (CA19142), with the aim of promoting access to media for people with disabilities in Europe. Book title: “Transforming Media Accessibility in Europe”, chapter title “A Multimodal Approach for Improving a Dialogue Agent for Therapeutic Sessions in Psychiatry” (25.09.2024).
3. Presentation of scientific results within the scope of the international Visegrad project (Visegrad +) called “AI Apps Against Alzheimer in Central Europe (4ACEs)” in Bratislava, Slovakia (17 – 18. 10. 2024).
4. Presentation of scientific results within the scope of the international Visegrad project (Visegrad +) called “AI Apps Against Alzheimer in Central Europe (4ACEs)” in Budapest, Hungary (20 - 21.01.2025).
5. Presentation of scientific results at the international conference “III International Digital Mental Health and Wellbeing Conference” in Grenada, Spain (21-23.05.2025).
6. Seminar talk on the topic of “Multimodal interaction with AI Agents” for the Summer School organized by Ss.Cyril and Methodius University of Skopje, North Macedonia (23 - 27.06.2025).

## 1.4 Participation in research projects

I have been involved in the following national and international projects. The research presented in the dissertation refers to the first of these projects (“Terabot: A dialogue system for therapy poznawczo-behawioralnej”).

National:

1. **Terabot - system dialogowy do terapii poznawczo-behawioralnej** (SzIR-2, IDUB), duration: 01.01.2021 – 30.06.2023:

The goal was to develop a dialogue agent targeting emotion recognition and management in a group of patients on the schizophrenia spectrum (F20.0-F29.0). The dialogue agent was used for a pilot randomised clinical trial with patients at the Institute of Psychiatry and Neurology, Warsaw, Poland. The project was financed by the Warsaw University of Technology from the Fund of Initiative of Excellence of the Research University (IDUB).

2. **PayEye - Research and Development** (Works for business entities), duration: 18.11.2021 – 30.11.2021:

The project consisted of comprehensive testing of PayEye’s product - a payment terminal using the user’s biometric features. It included preparing a test plan for the device with identification based on iris features, supervising tests with a diverse group of volunteers under varying conditions of use, and conducting a subsequent cross-sectional evaluation of the results obtained. Funding institution: PayEye sp. z o.o.

International:

1. **4ACEs - AI Apps Against Alzheimer in Central Europe**, duration: 01.07.2024 – 31.12.2025: Visegrad Fund (Visegrad +):

The 4ACEs project brings together three leading institutions (Institute of Informatics, Slovak Academy of Sciences (UISAV), Bratislava, Slovakia; Budapest University of Technology and Economics (TMIT), Budapest, Hungary; Warsaw University of Technology, Warsaw, Poland) from the Visegrad 4 region with expertise in AI-based diagnosis and treatment of neurodegenerative diseases, and INNOFEIT (Skopje, North Macedonia, lead partner) with its expertise with innovation and development. The project will serve as a framework for exchanging knowledge through on-site visits, workshops, training schools, and conferences. Project financed by European institutions - Visegrad Fund - Visegrad + Grants.

2. **Digital Mental Health for Young People (YouthDMH)**, COST action (CA23153), duration: 24.10.2024 – 23.10.2028, active member of Working Groups 1 and 4:

YouthDMH focuses on the role that digital technologies can play in supporting the mental

health needs of young people. It considers how technology can help mental health from the point at which young people first become aware of the difficulties, through to the delivery of large-scale, evidence-based interventions. The core aim is to maximize the positive application of technology while mitigating the potential negative impact. COST receives EU funding under the various Research and Innovation Framework Programmes, such as Horizon 2020 and Horizon Europe.

3. **Leading Platform for European Citizens, Industries, Academia and Policymakers in Media Accessibility (LEAD-ME)**, COST action (CA19142), 13.10.2020 – 12.10.2024, active member of all Working Groups:

LEAD-ME will boost cultural change and create a new mindset when designing tools for professional and private activities for European citizens of all abilities and disabilities. COST receives EU funding under the various Research and Innovation Framework Programmes, such as Horizon 2020 and Horizon Europe.

## 1.5 Other achievements

During my PhD, I completed a research stay, participated in various academic activities, and was recognized with multiple awards:

1. Third-Class Team Rector's Award, Warsaw University of Technology, for Scientific Achievements in 2022 - 2023" (2024);
2. Research Internship at the Technical University of Berlin (TUB), Germany - main focus: conducting interdisciplinary research in the field of Human-Robot Interaction under the supervision of Prof. Linda Onnasch, Professor of Psychology of Action and Automation (May - June 2023);
3. Cooperation with the University of Warsaw in conducting laboratory classes, course: "Biometric Identification of Identity" for Students majoring in Forensic Science (2022 - 2023);
4. Finalist in the Three Minute Thesis (3MT) competition organized at the Warsaw University of Technology to popularize science. Speech topic: "Where psychiatry meets IT - a multimodal dialogue agent for patient support" (December 2024);
5. Participation in the International Winter Training School on the topic of "Accessible Embodied Interaction" in Lisbon, Portugal (16 - 17.03.2023);
6. Participation in the International Summer School on the topic "Cognitive Engineering" at the Technical University in Berlin (Technische Universität Berlin, 4 - 15.07.2022);
7. Winning 1st place nationally at the 15th National German Language Olympiad for Students of Technical Universities, presentation on "Internet of Things - Möglichkeiten, Perspektiven, Zeitverwaltung" (2020).

# Chapter 2

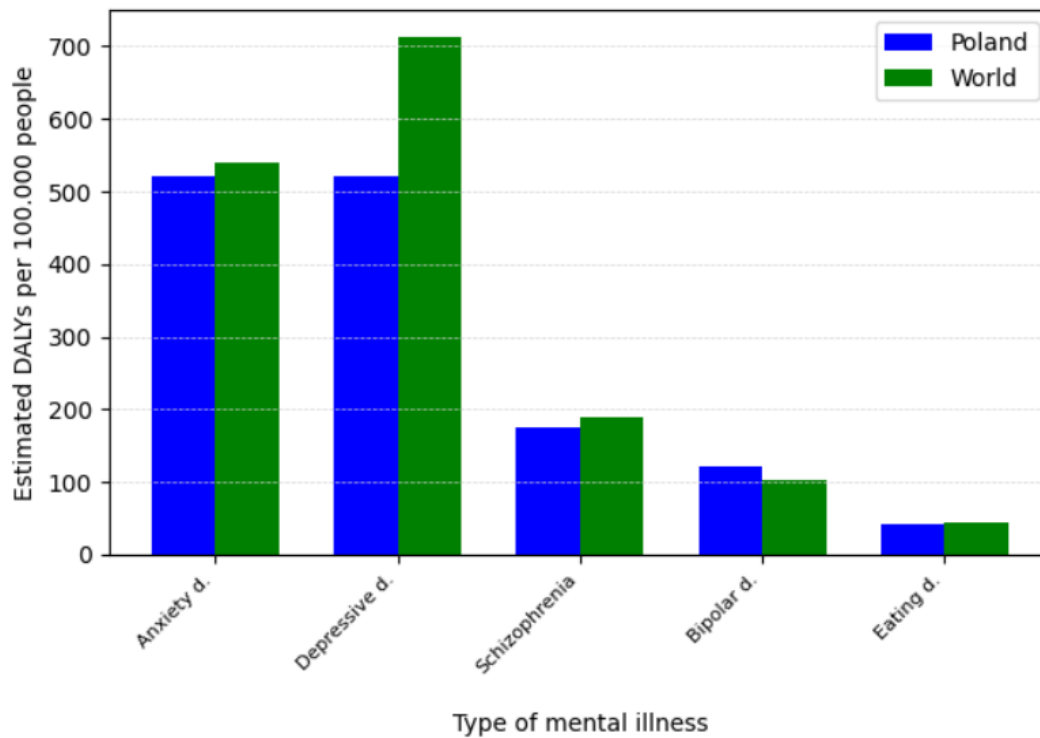
## Literature review

### 2.1 Mental health care as a global problem

Mental health problems can affect anyone and can have serious consequences in many areas of life. It is known that mental illnesses are determinants of work role disability and quality of life. In many cases, the impact of mental disorders is greater than that of common chronic physical disorders [7]. Studies from Harvard Medical School, based on the American population, indicate that approximately half of all Americans will experience a mental health condition that meets the criteria in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) at some point in their life [8]. It has been indicated by comparable European studies that mental health issues are also quite common, with a mental disorder being reported by around 25% of people at some point in their lives (also based on DSM-IV criteria) [9].

Figure 2 shows the burden of mental illnesses, based on data collected from the whole world and for Poland specifically. According to [11], the burden of disease is the largest and most comprehensive effort to quantify health loss across places and over time. This parameter is described as the impact of a health problem on a given population. It can be measured using various indicators, which allow it to be compared between different areas (countries, regions, etc.) It also enables future healthcare needs to be predicted [12]. It is a measure of the impact of living with illness and injury and dying prematurely, defined in disability-adjusted life years (DALYs) per 100,000 people. Figure 2 shows that depressive and anxiety disorders are the two leading mental health illnesses worldwide. These are followed by schizophrenia and bipolar disorders. The numbers may differ regarding Poland-specific data (from the year 2021), but still, again, anxiety and depressive disorders are leading among mental illnesses, followed by schizophrenia and bipolar disorders.

When it comes to children and young people, researchers concluded that the number of people affected by depression and anxiety disorders had increased over the years between 1990 and 2021 [1]. This increase accelerated significantly between 2014 and 2021. Although the growth rate of male depression was steeper, the burden was consistently higher in females.



**Figure 2:** Burden of mental health diseases (2021) worldwide and in Poland [10], where d – disorder

The fastest increases in depression and anxiety were observed in the 10–14 and 20–24 age groups, respectively [1]. By 2022, the percentage of adults in European countries at risk of depression was about 55%. The highest numbers were found in Poland, Greece, and Cyprus, where approximately 65% of adults were affected [13].

According to the World Health Organization (WHO), the state of mental health services is a significant problem. These services are often underfunded and poorly developed. In many countries, people with severe mental health conditions are unable to access the care they need. The estimation is that there are regions where up to 90% of them receive no treatment at all [14]. In March 2025, the WHO launched new guidelines to help countries reform and strengthen their mental health systems. These guidelines provide a framework for setting up modern, community-based services supported by the latest research and aligned with international human rights standards. The goal is to provide accessible, high-quality, respectful mental health-care to all [15].

A WHO report states that around 24 million people worldwide (1 in 222 people worldwide) are living with schizophrenia, experiencing symptoms such as hallucinations, delusions, and distortions of speech and thought [16]. It is also estimated to be the 10th most widespread non-fatal disease worldwide [17, 18]. As this is an incurable condition, patients require ongoing management through medication and psychotherapy. However, due to the associated costs and the insufficient number of specialists, there is a need to develop additional methods to support these patients and improve their mental well-being.

## 2.2 Dialogue (conversational) agents

In the literature on dialogue systems, the terms “dialogue agent”, “conversational agent”, “dialogue system”, and “chatbot” are used interchangeably. They are all described as a system capable of engaging in human-like social conversations. Some research papers on the system’s architecture, functionality, or purpose try to highlight their differences. An example of one definition [19, 20] is that chatbots are described as machine agents that serve as natural language user interfaces for data and service providers. It is explained that chatbots are typically designed and developed for mobile messaging applications. Different from chatbots that communicate via text, spoken dialogue systems have to be highlighted. They are developed to carry out conversations with users using spoken language [21]. Generally, one can state that all these phrases refer to software systems that mimic human dialogue with users [22]. Sometimes conversational agents (CAs) are also referred to as virtual agents [23, 24], or digital assistants [25]. A few examples of these assistants that are well-known from the industry include:

- Siri by Apple<sup>1</sup>,
- Alexa by Amazon<sup>2</sup>,
- Google’s Assistant<sup>3</sup>,
- Bixby by Samsung<sup>4</sup>,
- Celia by Huawei<sup>5</sup>.

In recent years, research in computational linguistics, multimodal interfaces, computer graphics, and autonomous agents has led to the development of even more advanced dialogue systems: Embodied Conversational Agents (ECAs). These agents can mimic human behavior in face-to-face conversations. One way in which this is demonstrated is through their ability to recognize verbal and non-verbal signals. [26]. These sophisticated systems have been designed and built for a wide range of different purposes, some of which include: banking, client services, human resources management, education [27]. Advances in natural language processing (NLP), speech recognition, machine learning, and AI have made conversational agents increasingly available. Additionally, dialogue systems have also been applied in the field of medicine and digital mental health with a variety of approaches.

Depending on their aim, two types of dialogue systems are described in the literature, mainly **non-task-oriented** dialogue systems and **task-oriented** systems. The non-task-oriented systems aim to provide conversations with the user on open domains with no specific task to

---

<sup>1</sup><https://www.apple.com/siri/>, accessed July 16, 2025.

<sup>2</sup><https://www.amazon.com/Meet-the-new-Alexa/dp/B0DCCNHV5>, accessed July 16, 2025.

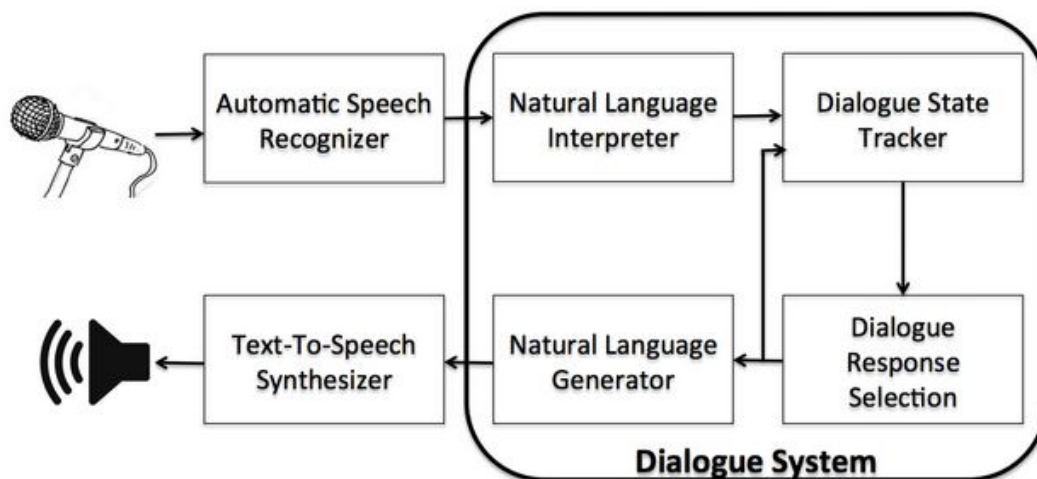
<sup>3</sup><https://assistant.google.com/>, accessed July 16, 2025.

<sup>4</sup><https://www.samsung.com/us/apps/bixby/>, accessed July 16, 2025.

<sup>5</sup><https://consumer.huawei.com/en/emui/celia/>, accessed July 16, 2025.

accomplish. Such systems are designed to mimic the long, unstructured conversations that characterise human-human interaction [28]. In the studies [29, 30], the authors explain that non-task-oriented systems are implemented using either generative or retrieval-based methods. Retrieval-based methods use response selection algorithms to select an appropriate response from a repository of answers for the current conversation. In contrast, generative models can generate more correct answers than could ever have appeared in the corpus. The task-oriented systems are designed to complete specific user tasks. Typical examples of their aims could be finding products, booking accommodation, or restaurants [30]. Today’s commercial systems can sometimes incorporate aspects of both of these approaches. Industrial chatbots (e.g. ChatGPT) can engage in longer, unstructured conversations, while digital assistants (e.g. Siri, Alexa) are typically frame-based dialogue systems [28]. This dissertation includes analyses of dialogues with Terabot, a goal-oriented spoken dialogue system designed for use in psychiatry. It is explained in detail in Chapter 3.1.

### 2.2.1 Main components of spoken dialogue systems



**Figure 3:** Illustration of a basic structure of a speech-to-speech dialogue system proposed by Serban et al. [23]

A key aspect of developing conversational systems is designing a dialogue management component that enables intelligent and engaging conversations [31]. One of the most common approaches to these systems is to treat them as a pipeline, which processes the inputs (words) of the user. Figure 3 shows an example of how a spoken dialogue system is composed [23]. In the case of speech-to-speech dialogue systems, the first component is an Automatic Speech Recognizer (ASR), then comes the Dialogue System itself, and finally a Text-to-Speech Synthesizer (TTS). Other systems (which are only text-based) do not have the speech synthesis components. Each component is a vast area of computer science with many different scientific approaches.

For this reason, only a brief description will be given in this dissertation. The aim of an ASR system is to take acoustic input from the user (via a microphone), to analyse it (usually using a pattern/model/algorithm), and to give an output (typically text) [32]. In simple words, an ASR converts recorded/detected voice into written text. In principle, it can be said that ASR systems use a combination of signal processing techniques to recognize the speech signal. The next step is then applying machine learning techniques and statistical modeling; they are used to match the speech signal to the corresponding words and phrases. The result is the production of a text representation of the spoken audio. Different classifications can be made when it comes to errors in ASR systems, which may significantly impact the behavior of a dialogue system or the dialogue flow. As the dissertation topic concerns dialogue systems, I have decided to present the following distinction, which might be relevant to the subject. Authors Halverson et al. [33] propose this differentiation in errors in ASR systems: direct, indirect, and intent. Direct errors are those that occur when a person misspeaks or stutters. Indirect errors arise when an ASR system misinterprets the speaker's input. Intent errors, on the other hand, happen when the speaker decides to repeat what has just been said. Apart from implementing this ASR approach, different services can be used in practice. The most commonly used are the solutions offered by Google Cloud Speech-to-Text, OpenAI Whisper API, and Microsoft Azure Speech Service. Another key area for dialogue systems is intention detection, since it is important to correctly identify the user's goal or motivation. The aim of TTS modules is to convert strings of text characters into spoken output. In practice, a variety of services can be used to implement a TTS approach.

The main core, meaning the **Dialogue System** itself, has several parts all connected to the Natural Language domain, which are:

1. Natural Language Interpreter: the role is to interpret the received utterance provided by the user and parse it into the next module;
2. Dialogue State Tracker: it aims to manage the history of the input and the dialogue and to define the current state of the dialogue;
3. Dialogue Response Selection: depending on the previous module, it is responsible for the selection of the response;
4. Natural Language Generator: based on the previously gathered information, the aim is to generate a valid response.

Figure 3 shows the main components of a spoken dialogue system, which was proposed in 2017 (before the development of large language model (LLM) based dialogue systems). Using LLMs is a more modern approach that will be described in Chapter 2.2.3.

## 2.2.2 Dialogue systems applied in the domain of digital mental health

To solve problems in digital mental health, researchers are trying to integrate conversational agents into this domain. In the last few decades, there has been much evidence of the potential benefits of using Embodied conversational agents (ECAs) for healthcare purposes. Due to the expanding possibilities of using Conversational Agents they have started to gain attention, especially when it comes to the topic of mental health and psychiatric diseases. CAs are relatively cost-efficient and easily accessible. Such a solution has been explored for many years with different approaches, with the main aim of improving the health of people struggling with mental health problems or diagnosed with mental illness [18]. Depending on the purpose of the CA, they have different target users or aims, e.g., assisting clinicians during therapies (conversations), or assisting patients and/or caregivers of older people in their everyday life [34, 35]. There are many scientific reports on different areas where CAs can be successfully used to benefit patients. In one of the existing review articles [6], AI-based CAs are described from other countries. A summarization of the findings shows that by 2023, there were 35 existing CAs connected to the digital mental health topic. They (and others developed by now) can be divided into different categories depending on:

- the **target condition of the users/patients** – e.g., broad (psychological well-being, anxiety, problem distress, diabetes-related distress, emotional support) but also some in particular, for example, breast cancer, PTSD, etc.;
- the **interaction mode** of the CA – is it text-based, voice-based, multimodal;
- the **delivery platform** of the CA – mobile app, website, robot, VR etc.;
- the **target users** –mothers, employees, students, diagnosed patients, etc.;
- the **therapeutic approach** provided – CBT, mindfulness, self-compassion therapy, person-centered therapy, etc.
- others – e.g. target country, study type.

Another category would be the description of the core of the CA (how the responses are generated). For example, is it rule-based, including neural networks, artificial intelligence, or large language models? Despite the different architectures, target groups, and handled topics, there is no doubt that conversational agents can provide an engaging way of getting psychological support at any time when it is needed [36]. Numerous conversational agents have been developed in the digital mental health domain, both in research and commercially. However, only a selected subset will be discussed in this dissertation. Following are several examples along with the relevant information and available details.

## ELIZA

One of the first CAs used for psychological therapies is called **ELIZA** [37]. Joseph Weizenbaum at MIT developed this text-to-text chatbot in the 1960s. It was an early NLP program to simulate a conversation between a human and a computer. Back then, the conversations between users and ELIZA had to be done with the help of a typewriter. When it comes to the core of ELIZA, it worked through several key mechanisms, e.g., keyword identification with ranking, text transformation based on rules, and fallback mechanisms for when no keywords were found. The paper stated that the program provided conversations as a person-centered psychotherapist (Rogerian psychotherapist) would do. This approach had a significant advantage: there was no need to store information directly related to the real world (e.g., background information about boats when the user was talking about going for a boat ride). ELIZA was seen as helpful since the author mentioned “it has an important psychological utility in that it serves the speaker to maintain his sense of being heard and understood.” [37]. Following this work, more and more research has been done on conversational agents with psychological uses.

## Woebot

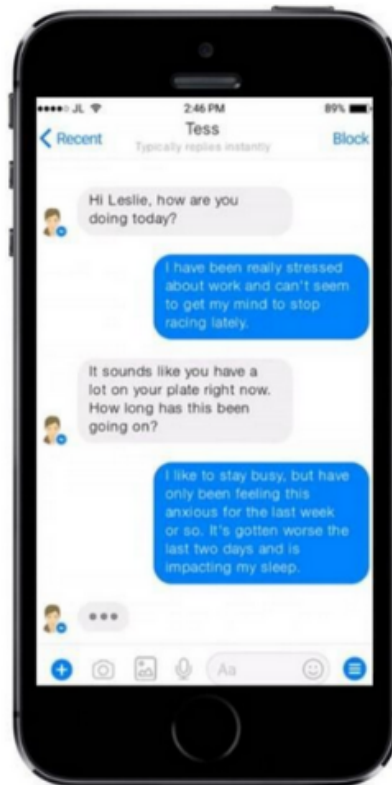


**Figure 4:** Snapshots of the Woebot conversational agent (chatbot) on a mobile phone [38]

Another conversational agent worth discussing is **Woebot**, which is shown in Figure 4. In this case, it was successfully employed for young adults in therapy for anxiety and depression [36]. This conversational agent was non-embodied, text-based, and provided dialogues in English. Regarding the core of Woebot, the authors claim that different computational methods were used depending on the section or feature. The main method used was a decision tree with suggested responses for Woebot’s utterances. Also, NLP techniques were embedded at specific

points in the tree to guide the conversations to the appropriate next steps [36]. The research using Woebot was conducted on young adults, students of Stanford University, who self-identified as suffering from symptoms of anxiety and depression. Then, based on interacting with Woebot for two weeks, they conversed with Woebot. The dialogues with Woebot were conducted using an instant messenger application on a mobile device or desktop application. Each conversation between Woebot and the user was about the user's general well-being on that day and the participant's mood. After collecting these data, CBT concepts were presented to the participant as short videos or simple word games. The aim was to explain cognitive distortions that might occur. The results of the study showed that, because of these conversations, the symptoms of depression and anxiety could be reduced. Woebot is a conversational agent that delivers elements of CBT engagingly and effectively to the participants. It gained many successes (both scientific and commercial), e.g., receiving 'FDA Breakthrough Device Designation' for its postpartum depression digital therapeutic, WB001. Until recently, this application could be used for free by every Messenger user, but a decision has been made that as of June 30, 2025, the chatbot will be retired by the company Woebot Health [39].

## Tess



**Figure 5:** A snapshot of the Tess conversational agent (chatbot) on a mobile phone [40]

Another conversational agent is called **Tess**, as shown in Figure 5. The authors claim that it is AI-based and offers cognitive behavioral therapy elements to its users [41]. It was designed

to relieve self-identified symptoms of depression and anxiety in college students. The study reports that 75 participants were recruited from 15 universities across the USA. Designed by the company X2AI Inc., Tess provides brief conversations through integrated support, psychoeducation, and reminders. Tess is text-based; it can be integrated into existing communication applications, e.g., Facebook Messenger and Slack, but only via SMS. Tess is described as an adjustable platform that allows content to be altered to a specific form of the user's treatment or demographics. For instance, it has been adapted to deliver interventions to reinforce weight management goals in adults with pre-diabetic symptoms [42]. The results show that students who used Tess had experienced a significant reduction in depression and anxiety [43].

## Wysa



**Figure 6:** A snapshot of the Wysa conversational agent (chatbot) on a mobile phone [44]

The next conversational agent described here is **Wysa** (a snapshot is presented in Figure 6). This text-based chatbot is available in a mobile application (available for the Google Play Store and the Apple App Store). It comes with other supporting functionalities that the app provides (e.g., booking a session with a real psychotherapist, taking part in relaxation exercises, getting information about emotions, and sleeping habits). Wysa aims to reduce symptoms of depression and anxiety (helps users with self-reported symptoms of depression based on the Patient Health Questionnaire – 9). Additionally, it helps to minimize pain interference and improve physical function. It is important to mention that the chatbot is always available and free. The authors claim that Wysa utilizes AI elements, but no other information is provided on what exactly is

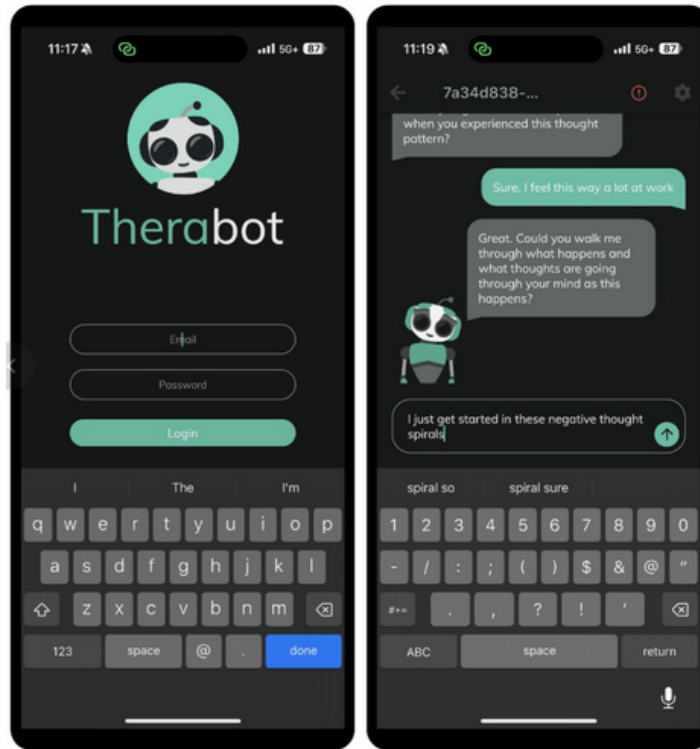
used [45]. The study analyzes high-engaged and low-engaged users by analyzing in-app user feedback. The study outputs are promising. Findings show that the users with high engagement with the application experienced a higher average improvement in their well-being compared with the low engagement users group. It is worth mentioning that in 2022, the US Food and Drug Administration (FDA) granted a Breakthrough Device Designation for Wysa's AI-based digital mental health conversational agent [46].

### **Fido**

A fairly recent article [47] describes a Polish-language text-based chatbot named **Fido**. This chatbot was integrated into Facebook Messenger, as the authors explain. The selected participants were added as testers to Meta's development website, where they could access the chatbot via links sent in individual emails. Fido provides psychoeducation about depression and anxiety. It also suggests gratitude exercises. The description of the core of this chatbot states that intent recognition was possible using machine learning models (no further details are given). Fido was implemented by Emplocity and runs on NLP technology provided by this company. The study shows that it was tested on people with anxiety and depressive behavior. In terms of results, the authors claim that exploratory analysis showed a decrease in loneliness among participants who used Fido more frequently. They also showed that participants' depressive and anxiety symptoms were reduced.

### **SchizoBot**

Finally, **SchizoBot** represents a chatbot specially designed for patients with schizophrenia [18], but has not yet been tested on patients. SchizoBot is described as a text-based and AI-based mental health application to provide CBT therapy and support to patients diagnosed with schizophrenia by chatting with them. The study received ethical approval in 2022. The authors claim that it uses artificial neural networks (ANN) to analyze collected data and predict the outcome of SchizoBot's utterances. In the architecture, elements of machine learning techniques and deep learning methodologies were incorporated. The dataset of this conversational agent was collected from journals from PubMed, questions posed on the NCBI website by patients diagnosed with schizophrenia, and the ISEAR dataset (which contained data on seven emotions). The authors further processed the data; their findings describe the ANN architecture, the training, and hyperparameter optimization in detail in their findings [18]. This conversational agent is a retrieval-based model with predefined patterns and responses, offering personalized CBT conversations. The paper does not contain information on whether SchizoBot has been tested on patients with schizophrenia or evaluated by them in any way. Despite that, it shows promising implications for the domain of digital mental health.



**Figure 7:** A snapshot of the Therabot conversational agent (chatbot) on a mobile phone [48]

## **Therabot**

A recent report shows another (text-to-text) chatbot, named **Therabot** [49]. Figure 7 shows a sample dialogue of this chatbot. The authors claim that it has a generative AI expert-fine-tuned component for dialogue management. The chatbot was tested in a randomized, controlled trial with 210 adults having symptoms of depression, anxiety, or a risk of eating disorders. The participants took part in either a four-week intervention (the group that interacted with the chatbot) or were assigned to a waitlist control; they received no app access during the study period but gained access afterwards. As the results show, the Therabot users experienced reductions in all of their symptoms (depressive, anxiety, and eating risk disorders). They used Therabot on average for more than six hours during the study. The participants rated the therapeutic engagement as comparable to that of human therapists. There is a high chance that more reports on conversational agents, especially using AI or generative AI components, will be published in the near future. This indicates that there is a large research gap in the field of digital mental health, and at the same time, there is a high demand for such technological solutions in society.

### **2.2.3 LLM-based solutions for psychological well-being and mental health**

Large language models (LLMs) are a class of large-scale artificial intelligence models, with billions of parameters [50, 51]. These LLMs can learn complex and accurate language representations. This improves their performance in NLP, particularly in natural language generation

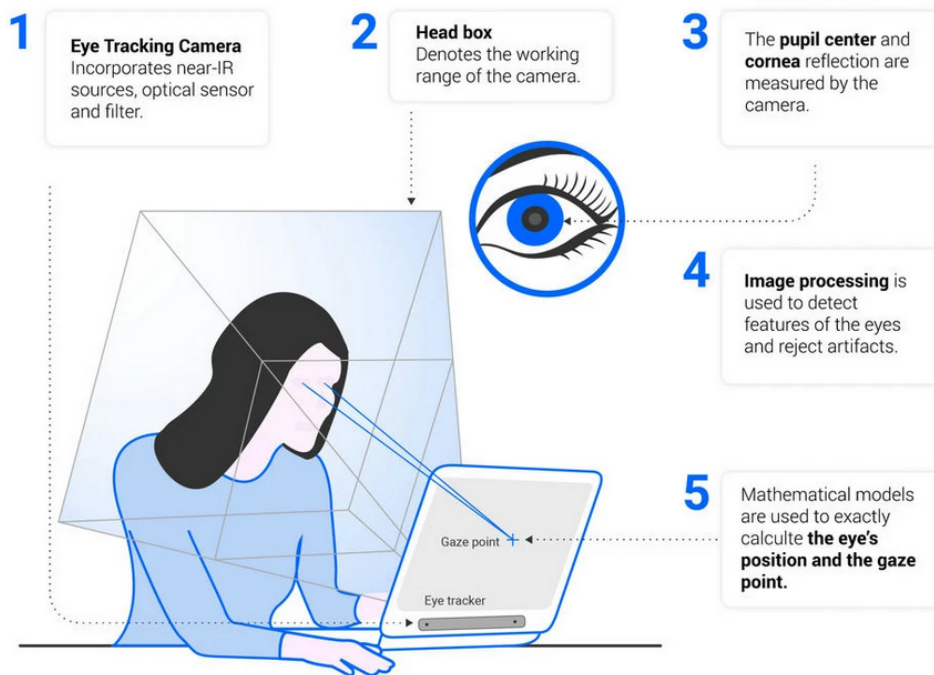
tasks [50,52]. The development of generative AI in recent years has opened up new possibilities for developing mental health solutions, especially regarding dialogue systems. LLM-based conversational agents and chatbots are gaining more and more attention in mental health care. They are likely to be better at handling complex information – they can give personalized, flexible, and helpful responses to specific mental health needs [6].

Studies have already been conducted on how to make LLMs suitable for a variety of tasks in the field of psychological well-being. For example, studies describe how LLMs have been used to guide people in making important decisions [53] or providing relationship advice [54]. One such example is Socrates – a personalized chatbot that has been developed on the ChatGPT platform. Fine-tuned to support psychological well-being, it promotes open and self-reflective conversations [55]. The research paper explains that the system’s functionality draws upon a carefully selected repository of knowledge from psychological, psychotherapeutic, and philosophical texts, which is integrated into its design. Research has been carried out into the effectiveness of LLMs in providing mental health support. One example of this is when responses by LLMs were evaluated against human responses, it was found that they demonstrated more empathy and helpfulness when responding to relationship or general health-related questions [56]. In recent research, a study also describes when GPT-4 was prompted to act as a therapist. In this case, it demonstrated competence, empathy, and therapeutic capacity when providing one-on-one therapy to individuals seeking assistance with relationship problems [54]. There are already comparisons published discussing different mental health chatbots (based on LLMs) that differ not only by category (e.g., depression, anxiety, stress, suicidal thoughts), but also by target groups (e.g., mothers, college students, workplace employees) [57]. These approaches indicate the potential for using LLMs to assist human therapists in their work.

Despite the benefits of using LLM-based dialogue agents, researchers must be aware of the probability of giving an inappropriate response to the user. The risks of LLMs used in chatbots for therapeutic purposes should be considered. These chatbots should be designed with regard to crucial mental health guidelines. According to the magazine “The Brussels Times” [58], in March 2023, a Belgian man committed suicide as a result of following exchanges with a chatbot. He was faced with eco-anxiety, so he reached out to talk to an LLM. It was a chatbot powered by GPT-J, an open-source artificial intelligence language model developed by EleutherAI. Following six weeks of in-depth discussions, he ended his own life. Mathieu Michel, Secretary of State for Digitalization, commented on this with the following statement: “With the popularization of ChatGPT, the general public has discovered the potential of artificial intelligence in our lives like never before. While the possibilities are endless, the danger of using it is also a reality that has to be considered”. This incident clearly shows an urgent need for researchers to investigate and further develop LLMs’ safety guidelines, especially for mental health purposes.

## 2.3 Eye tracking

Eye tracking is a technology used to observe and analyse how a person looks at an object. The aim of measuring and analyzing eye movements is to gain insight into the attentional behavior of the viewer. Additionally, the observer's visual attention path can be followed [59, 60]. Eye trackers use infrared light technology with a high-resolution camera (or other optical sensor). Most of them use the near-infrared (NIR) light spectrum with the light wavelength  $\lambda$  in the range of  $0.8 - 2.5 \mu\text{m}$  [61].

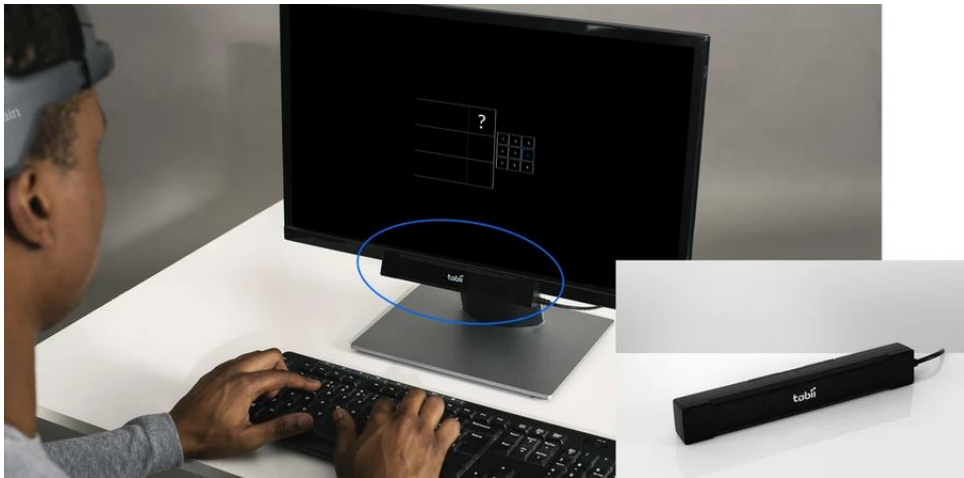


**Figure 8:** Schematic representation of the eye tracker workflow [62]

It is important to note that most commercial eye-tracking devices are based on an infrared illumination approach. However, there is also research into appearance-based systems that use **webcams** instead of infrared light to extract and track eye features on the detected face area of the user. This kind of eye-tracking approach is often used in remote testing (from the participant's home or work). This approach comes with a big disadvantage: poorer accuracy than commercial near-infrared eye trackers. In a remote session, participants often move their heads and/or their computers, which can hardly be controlled and decreases the accuracy of webcam-based eye tracking. What is also important to mention is that the sampling rate used in the webcam-based approach for gaze location is relatively low, greatly influencing the data analysis that can be performed afterwards [63].

- Eye-tracking devices – design

There are several different types of eye-tracking devices commonly produced in terms of their design. Most of them can be divided into three main groups [68]:



**Figure 9:** Example of a stationary eye tracker model: Tobii Pro X2-30 by Tobii [64]



**Figure 10:** Example of a stationary eye tracker, model: EyeLink 1000 Plus by EyeLink [65,66]



**Figure 11:** Example of eye-tracking glasses (mobile eye tracker), model: Pupil Neon by Pupil Labs [67]

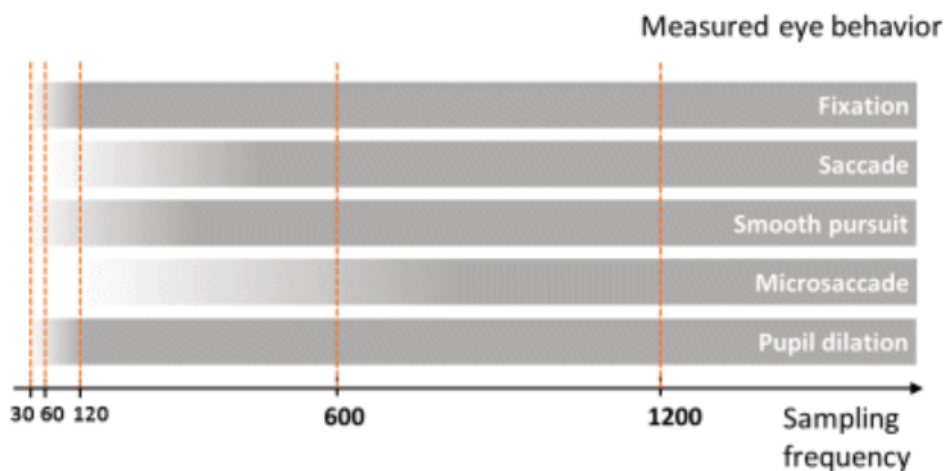


**Figure 12:** Example of an integrated eye tracker into an AR/VR system: Pupil Labs integrated into HIVE VIVE VR system [65]

- **stationary** eye trackers, also called remote or screen-based eye trackers (see Figure 9 and Figure 10);
- **mobile** eye trackers, also called wearable or eye-tracking glasses (see Figure 11);
- **integrated or embedded systems** in augmented/virtual reality devices (see Figure 12).

The main difference between the three is where the equipment is located during eye tracking. Wearable eye trackers, or eye-tracking glasses, are worn on the participant’s head. The same applies to integrated eye tracker devices in augmented/virtual reality devices. On the other hand, stationary eye trackers are entirely non-invasive and contact-free. They are positioned at a certain distance from the user. These distances are specified in the device manual (for the Gazepoint GP3, for example, this is approximately 50–100 cm). From a scientific and practical perspective, the choice depends on the purpose and features of the research environment in which the eye tracker will be used.

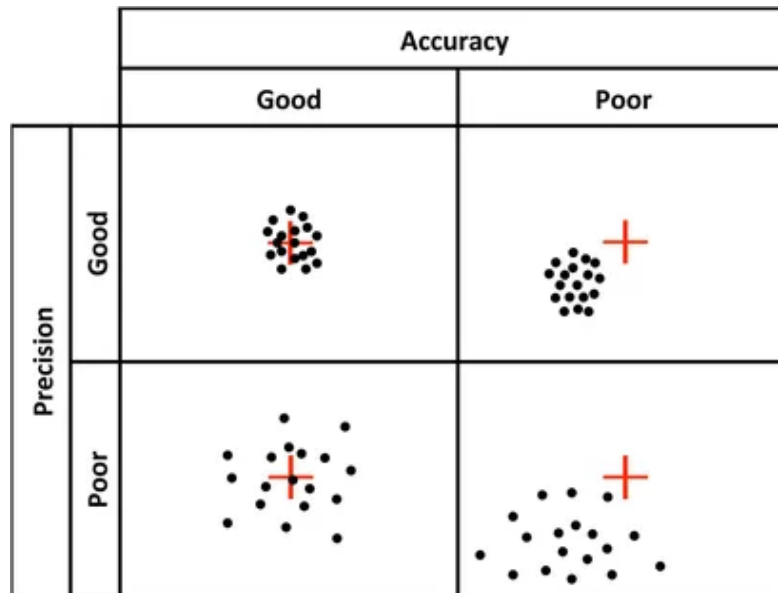
### 2.3.1 Technical parameters of eye trackers



**Figure 13:** Sampling frequencies and the corresponding data parameter measurement [69]

The **sampling rate** is considered to be one of the most important features of an eye tracker (measured in Hertz [Hz]). It is the number of samples of a specific parameter collected during one second. According to [63] and looking at available devices offered by different eye tracker companies, the sampling rate varies from 25 - 2000 Hz. The purpose of the eye tracker (or the research study in which a specific device will be used) predefines the sampling frequency that should be chosen. One of the producers of eye-tracking devices, Tobii, proposed an easily interpretable graph (presented in Figure 13). The graph indicated the parameters that can be collected during an eye-tracking session and the corresponding sampling frequency [69].

In general, it is considered that sampling rates above 250 Hz are required for research measuring saccades (e.g., their speed) [63]. An older research article [70] reports that a sampling frequency of at least 300 - 400 Hz allows the most accurate evaluation of the maximum velocity of saccades. Saccades can be of interest to research fields like medicine, e.g., neurology, for early detection of neurodegenerative diseases (e.g. [71,72]).



**Figure 14:** Illustration of accuracy and precision of an eye tracker [73]

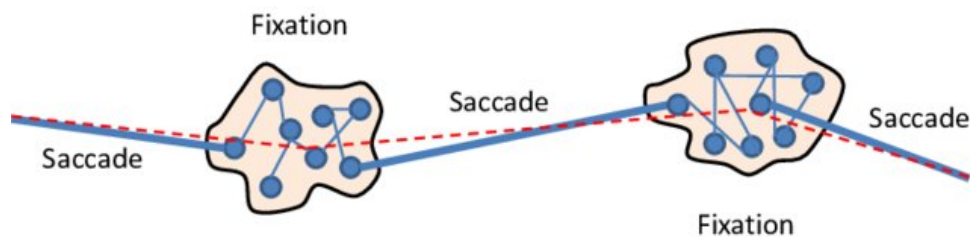
Regarding the quality of the data an eye tracker collects, two critical parameters should be discussed: accuracy and precision. **Accuracy** can be described as the average difference between the recorded and actual fixation positions. This is measured in degrees. Ideally, this parameter would be as small as possible, but this is not practically achievable. According to [63], this parameter is usually between 0.5 and 1 degree. This one degree corresponds to approximately 1.3 cm on a computer screen when the distance between the eye tracker and the screen is about 69 cm. Other research papers (e.g., [74]) also report that, particularly for remote eye trackers, the difference between the actual gaze point and the measured point is often larger than one degree, even in controlled environments. It is also important to note that the accuracy values described in different eye tracker manuals are measured under ideal conditions, meaning that the participant does not wear glasses and has performed the calibration ideally before using the device. It should be noted that these perfect conditions are not always feasible in practical applications involving real-life experiments. This is why the actual accuracy error may be slightly higher.

The **precision** of an eye-tracking device is a parameter that indicates how accurately a specific measurement can be reproduced. In an ideal scenario, the eye tracker should report the exact location of the user's eye in two consecutive measurements, assuming the user's eye remains in the same position. Once more, this cannot be achieved, for example, due to the human eye's nature (micro eye movements). Precision values for eye tracker devices range from 0.01 to

1 degrees. It results from the root mean square of the distance between two subsequent samples. However, it should be noted that the precision values reported by eye tracker manufacturers in their manuals result from using a motionless artificial eye. Therefore, in practice, the precision value will also be higher than reported. Figure 14 summarizes these two parameters and may give a better indication.

### Eye-tracking data including saccades and fixations

The authors Just, Carpenter et al. [75] explained that eye tracking is based on an *eye-mind* hypothesis. They claim that the location of a fixation indicates the area of interest (when looking at a visual display and performing a task). Other researchers describe this hypothesis in different words, meaning that gaze position is typically associated with what a person is currently thinking about (especially when there is a goal in mind) and what this person is currently paying attention to [63]. Sequences of fixations and saccades guide human perception. This hypothesis indicates the two main parameters that can be collected during eye-tracking sessions: saccades and fixations.



**Figure 15:** Illustration of saccades and fixations and microsaccades [76]

**Saccades** are fast movements of the eyes occurring from one point of interest to another (which are fixation points). Their range in amplitude depends on the task: from small, e.g., while reading, to larger, e.g., when gazing around a room. Because of the rapid movement during a saccade, the image on the retina is of poor quality, so information is mainly taken during the fixation period [77]. Saccades last on average between 20 and 40 ms. They can be divided into those that are triggered voluntarily or involuntarily. The saccade latency can be defined as the time required to formulate a saccade plan, depending on the specific task (typically ranges from 100 to 1000 ms). Once the eye has started moving, the endpoint of a saccade cannot be changed. Blinks can be estimated indirectly, from data loss in the pupil diameter signal (providing information about blink rate and duration), or directly, from the eye openness signal. According to [77], saccades are the fastest movement the human body can produce.

**Fixations** occur when the eyes stop scanning the scene and focus on a specific area of the field of vision. This enables the brain to process detailed information about what is being observed. According to [77], fixations last between 50 and 600 ms. In practice, they are usually

analyzed by their location (i.e., what is being fixated on) and by computing their frequency and duration on an area of interest (AOI).

**Micro eye movements** occur because the eye is not entirely still during one fixation. These small displacements of the eyeballs ensure that vision does not fade during fixation. There are three classes of them: **tremor** (the smallest with low amplitude and high frequency; often overlaps with drift; their role is not yet clear), **drifts** (slow, irregular movements of the eye that occur during the epochs between microsaccades [78]); **microsaccades** (the largest; they quickly bring the eye back to its original position). Human perception is considered stable despite the presence of fixational eye movements [79].

**Blinks** are a natural reflex that keeps the eyes lubricated and protects them from irritants like dirt and dust. Blinking is also a form of protection against bright lights and potential injury. On average, most blinks last between 150 and 400 milliseconds. The eyes close fully for around 50 milliseconds during that blink. Blinking can be voluntary, involuntary, or reflexive (in response to external stimuli). According to [77], the frequency of blinks varies according to age:

- Fetuses: less than three blinks per minute,
- Children: six to eight blinks per minute,
- Adults: around 20 blinks per minute.

In practice, blink rate and duration can be estimated indirectly from data loss in the pupil diameter signal or the eye openness signal [80]. In general, three main groups of eye-tracking data can be collected in eye-tracking studies; these are shown in Table 2. Other parameters collected during eye-tracking studies include pupil diameter, number of blinks, blink duration, time to first fixation, and many others. It is important to note that the choice of a particular parameter to collect during a study and its analysis always depends on the purpose of the experiment.

Different research areas and topics require different parameters to focus on and analyze. The use of such eye-tracking devices applies in a variety of different areas, including usability, marketing, advertising, industrial engineering, human factors and ergonomics, neuroscience, psychology, psychiatry, and psycholinguistics, not to forget computer science [59]. They are useful for understanding reading, scanning, and visual stimuli processing through information collected from eye movements [81]. The capabilities and applications of eye tracking are vast and expanding as new applications are created and developed. Some brief examples are seen in psychology when analyzing emotions, the important parameter would be pupil size (when different stimuli are presented to the participant); in medicine (neurology), where eye tracking can be used for early diagnosis of psychiatric and neurodegenerative diseases. There the focus would be on saccades and related parameters [71] and micro eye movements (e.g., microsaccades [72]). Another example is the output data from eye tracking can be used as an input to algorithms that are an important part of the HCI domain. Chapter 2.3.2 describes these applications in more detail.

**Table 2:** Eye-tracking parameter groups and their description, based on Gazepoint GP3 manual

Parameter Group	Description
Eye Movements	Fixations: number, duration, X/Y-coordinates of the fixation on the screen (FPOGX/Y)
	Saccades: number, duration, direction, magnitude
Blinks	Identifier of a blink
	Duration of the blink (ms)
Pupils	Right pupil size (in mm or pixels) (RPMM)
	Quality flag of the right pupil – valid or not valid (RPV)
	Left pupil size (in mm or pixels) (LPMM)
	Quality flag of the left pupil – valid or not valid (LPV)

### 2.3.2 Eye tracking in Human–Computer–Interaction

For some years, eye trackers have also been used in human–computer interaction with ECAs being an example of this application. Contact is established with humans through speech and non-verbal behavior [82, 83]. Since dialogue systems need to become more trustworthy, these ECAs should be equipped with communicative and expressive abilities similar to those we know from human-to-human interaction (speech, gestures, facial expressions, gaze, etc.) [83]. The eye-tracking data collected from humans during such dialogues is then used to support or control conversational agents. Some examples of dialogue agents that take into account the participant’s eye-tracking data are:

- Gandalf, a humanoid agent that can narrate planets and moons [84];
- Emma, the interactive storyteller [83];
- a dialogue agent that interacts with seniors [85].

There are also studies about collecting eye-tracking parameters characterizing a person’s emotional state [86] and level of concentration [87]. Various approaches and enhancements in the quality of human-computer interaction aim to make it more natural. As a result, dialogue agents can interact more authentically with people. This is a clear example of how multimodal approaches in dialogue agents can create an engaging user experience. A special case of dialogue agents using eye-tracking data as additional input is attentive user interfaces (AUIs). They can be described as user interfaces sensitive to the user’s attention. These AUIs use extra channels of input that measure characteristics of the user’s attention (e.g., user presence, speech activity, gaze) [88]. These systems can determine the urgency and relevance of the information

or actions they offer (in the context of the user's current activity) [88]. What is important is the monitoring of the user's attention during the interaction with the system. Based on this information, the system can adapt and give context-based options to the user. This makes the interaction more suitable for the current user's behavior/information state. Some AUIs can, for example, gather information from the users' gaze behavior about their cognitive state [89]. In some research papers, eye tracking is considered the most relevant attentive input technique [88], so gaze is regarded as the primary input modality for attentive user interfaces [89]. Knowledge of the user's gaze position plays an important role for AUIs; gaze data can provide helpful information about the context of the user's action. The system can further use this information to actively adapt the user's information state/behavior. In the research paper [90] it is suggested: *"To design less intrusive and more sociable interfaces, we suggest augmenting computing devices with attention sensors that allow them to prioritize their demands for user attention."* The authors also introduced five key components of an appropriate AUI as a framework for augmenting user attention [91]:

- **Sensing attention:** AUIs can determine which device/person/task a user will most likely be engaged with by tracking the user's physical presence, body orientation, and eye fixations.
- **Reasoning about attention:** interfaces can estimate the priority of user tasks by statistically modeling simple interactive user behavior.
- **Communication of attention:** AUIs should provide information about the user's attention to other people/devices. Communication systems should provide information about who/what the user focuses on and whether the user is available for communication.
- **Gradual negotiation of turns:** it is advisable that interfaces should determine the availability of the user for interruption. Firstly, they should check the request priority, then they can signal the request on a peripheral channel. Ultimately, they should detect the user's acknowledgment of the request before taking the turn.
- **Augmentation of focus:** the ultimate goal of all AUIs is to increase the user's attention. For example, analogous to the cocktail party phenomenon, AUIs can enhance information that the user is focusing on and reduce surrounding details.

The authors also pointed out that when designing AUIs, it is important to treat user attention as a limited resource. In human conversations, people can only listen to and absorb one person's message at a time, as is well known from social psychology [92]. Attention should be seen as a central channel for interacting and developing social ways of communicating [91]. In the literature, a parallel is drawn between an AUI and the modern traffic light system [88,91]. With the great potential of AUIs, researchers have also pointed out the problems associated with this topic: the application should not react every time the target of the gaze changes, but only

in appropriate situations and at the right time. This is known in the literature as the “Midas touch” problem. Bad design of an AUI system can create interruptions that are intrusive and annoying when they do not follow reasonable conventions for control flow. While humans are good at recognizing when the gaze requires action, for systems, these situations can be a major challenge [93]. To the best of my knowledge, there is a research gap, where there is no evidence of a dialogue agent enhanced using eye tracking and conversing in the Polish language. This dissertation aims to fill in this gap.

# Chapter 3

## Terabot system

### 3.1 Terabot – a RASA-based dialogue system in the Polish language for patients with psychiatric disorders

Terabot, a therapeutic spoken dialogue system operating in the Polish language, was designed by a research team at the Warsaw University of Technology (WUT), Faculty of Electronics and Information Technology. This research was funded by the Center for Priority Research Area Artificial Intelligence and Robotics of the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) program. The aim of the dialogue system was to help patients diagnosed with psychiatric diseases deal with their emotions. It was goal-oriented, meaning the main task was to talk with the patients about one of the chosen emotions – anxiety, fear, or shame. A relaxation exercise was recommended for the patient at the end of the conversation, which was intended to help calm their emotions. All answers given by Terabot were checked and approved by a professional (psychiatrist and psychotherapist). Using elements of CBT, Terabot’s dialogue sequence followed psychiatric recommendations (meaning Terabot’s utterances were based on CBT). It was a digital tool that could be used as an addition to therapies for many mental disorders (e.g. depression, anxiety disorders, obsessive-compulsive disorders).

#### 3.1.1 User interface

Terabot, as a spoken dialogue system, was embodied using a film of a man wearing a hygiene mask. He moved slightly, and his eyes blinked naturally. This gave the patient the impression of sitting in front of the computer during an online session with a therapist. The user interface is presented in Figure 16. At the top of the screen, there was a statement in the Polish language, saying “*Powiedz ‘Cześć’, aby rozpocząć rozmowę z Terabotem*”, meaning “*Say ‘Hi’ to start a conversation with Terabot*”. This helped the patients to start the conversation whenever they felt ready to talk. At the bottom of the screen, there were only two available items - on the left,



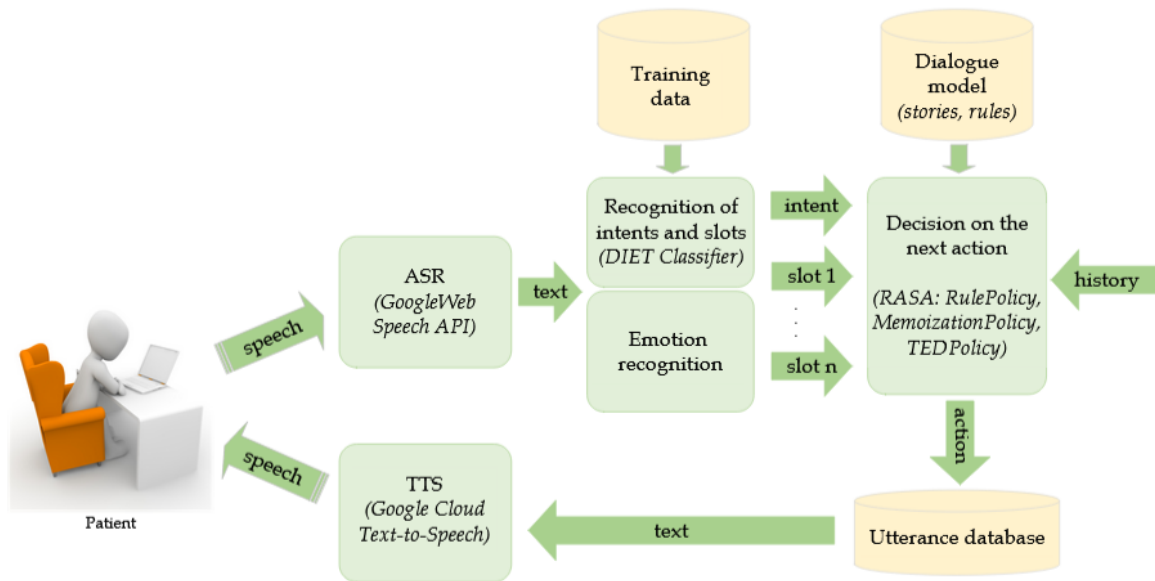
**Figure 16:** An image of Terabot's visual interface

the microphone symbol, and on the right, a 'Reset' button. The microphone symbol turned red, indicating that audio is being recorded in the moments when the patient should answer. When the patient should listen to Terabot's utterances, meaning it does not record audio anymore, the microphone turned black. The 'Reset' button was designed in case something goes wrong during the conversation, causing the need to restart.

### 3.1.2 Architecture

The dialogue system consists of several different modules listed below:

- Automatic speech recognition (ASR) module – in this case GoogleWeb Speech API was used;
- Dual Intent and Entity Transformer (DIET) classifier
- Text-based emotion recognition module
- RASA – an open-source framework for dialogue management and natural language understanding (NLU),
- Utterance database – a collection of statements for Terabot's utterances. Terabot's utterances were co-designed by a medical professional (psychiatrist and psychotherapist); thus, it was guaranteed that Terabot would give professional answers.
- Text-to-Speech (TTS) module – in this case the Google Cloud Text-to-Speech service was used.



**Figure 17:** Terabot dialogue system based on the RASA framework

When the patient starts speaking, the ASR module recognizes this speech signal and converts it into text. As previously mentioned, Google Web Speech API for the Polish language was used. This text is then passed to the Dual Intent and Entity Transformer (DIET) classifier. This is a multitask transformer that analyzes the received text. It can handle both intent and entity recognition. The DIET uses a sequence model that considers the sequence of words [94]. For example, the intents that are classified are: *greet, goodbye, affirm, deny, say mood, say story, choose exercise*, etc. Additionally, the slots are filled in with, for example, the exercise name or the recognized patient’s emotional state. There is also an emotion recognition module based on the Bidirectional Encoder Representations from Transformers (BERT) [95], and was part of previous work made for Terabot’s purpose [96]. It uses the BERT model with fine-tuning for emotion classification. At the same time as the intent and slot recognition, the output of the ASR is passed here. The currently detected emotional state of the patient’s spoken text determines the value of the “emotion” slot, and another slot is filled with the exercise type.

In the next step, a decision about the action of the following dialogue system has to be made. This decision is an output of three components, namely a weighted combination of the rule policy, a memoisation policy (i.e. based on stories stored in the memory), and a Transformer Embedding Dialogue (TED) policy [97]. This is shown in the right green rectangle of the diagram shown in Figure 17. The TED policy considers the current state of the dialogue, including the patient’s utterance intent and slot values. It is important to mention that the DIET classifier and the action decision pipeline are realized using RASA [98]. This is an open-source dialogue management and NLU framework. When the choice of the next system action is to give Terabot’s utterance, it is selected from the utterance database. In the next step, the suitable utterance is transformed from text into a speech signal using Google Cloud Text-to-Speech API

for the Polish language. The result is Terabot's response given to the patient.

## 3.2 Experiments at the Institute of Psychiatry and Neurology

This chapter provides an overview of the study conducted with patients diagnosed with schizophrenia at the Institute of Psychiatry and Neurology in Warsaw (IPIN). It was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee. It received approval on 27 April 2022 from the Ethics Committee of the Institute of Psychiatry and Neurology in Warsaw, Poland (resolution No. IV/2022). The study took place between March and August 2023. A total of 38 participants were selected for the study. All of them were admitted to a 24-hour psychiatric hospital. These patients were diagnosed with various types of schizophrenia, i.e., F20.0 to F20.9 according to ICD-10 (WHO, International Classification of Diseases [99]). Schizophrenia is a chronic brain disorder that causes a wide range of behaviors, which may include hallucinations, delusions, disorganized communication, cognitive difficulties, and decreased motivation. Most of the symptoms can be significantly improved with proper treatment, which also reduces the probability of relapse. Unfortunately, schizophrenia is still incurable, but ongoing research continues to develop more effective and safer treatment approaches [100, 101]. Below, fragments of the patient's behavior description, especially concerning speech, are presented (according to WHO [99]):

- F.20.0 – paranoid schizophrenia – “ (...) Disturbances of affect, volition, and speech, and catatonic symptoms, are either absent or relatively inconspicuous (...)”;
- F.20.1 – hebephrenic schizophrenia – “(...) behavior irresponsible and unpredictable, and mannerisms common. The mood is shallow and inappropriate, thought is disorganized, and speech is incoherent (...)”;
- F.20.5 – residual schizophrenia – “(...) passivity and lack of initiative; poverty of quantity or content of speech; poor nonverbal communication by facial expression, eye contact, voice modulation and posture (...)”.

Study participants were recruited from the Psychosis Relapse Prevention Ward at IPIN. This facility serves inpatients recovering from a mental health crisis. The program for these patients included a range of psychological interventions, such as therapy groups, metacognitive training, community meetings, psychoeducation, initiative workshops, physical activities, and cognitive training. An experienced psychiatrist admitted the patients according to the criteria presented in Table 3.

Based on their diagnosed conditions, the patients were treated with medication, including antipsychotics, mostly in combination with other antidepressants or mood stabilizers. They were examined and classified by psychiatrists for additional therapy sessions with Terabot.

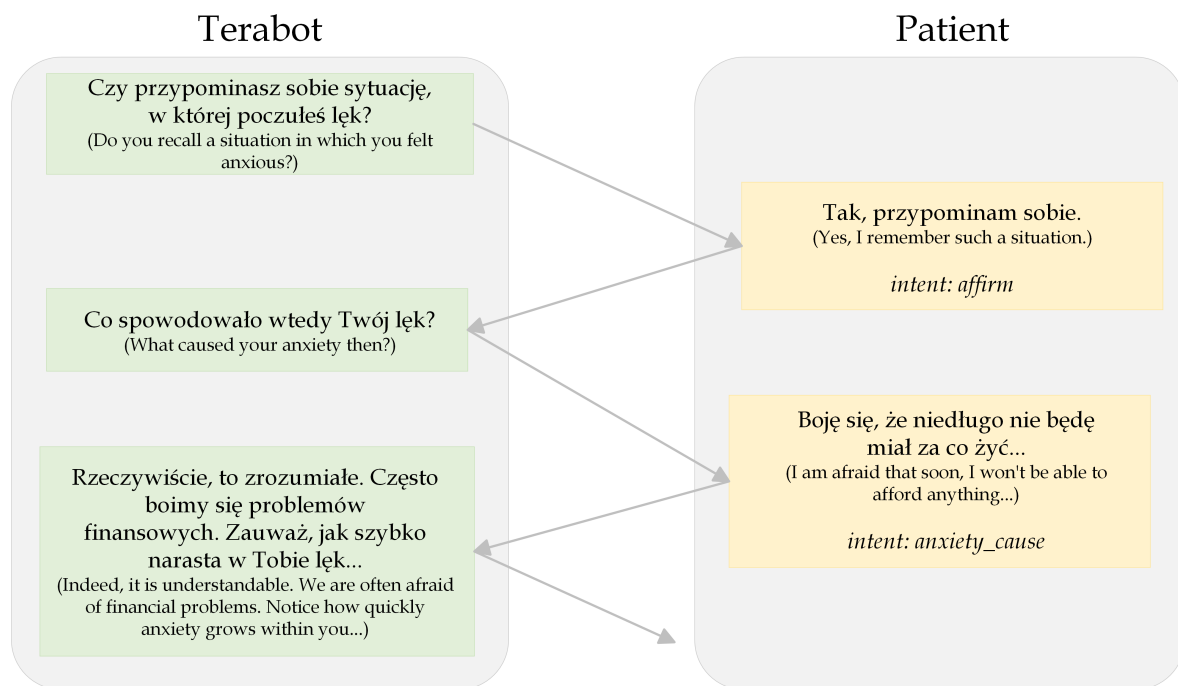
**Table 3:** Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
<ul style="list-style-type: none"><li>- age between 18 and 65,</li><li>- meets the diagnostic criteria for schizophrenia according to ICD-10 (F 20.0 - 20.9),</li><li>- no changes in prescribed pharmacotherapy in the week preceding the study and during the intervention,</li><li>- no documented mental disability or organic changes of the central nervous system,</li><li>- state of mental state stability allowing for the use of the dialogue agent,</li><li>- no coexisting active addiction to psychoactive substances,</li><li>- computer illiteracy.</li></ul>	<ul style="list-style-type: none"><li>- aged under 18 or over 65,</li><li>- refusal to participate in the study,</li><li>- acute psychotic crisis,</li><li>- the use of sedative medications the week preceding the study and during the intervention,</li><li>- cognitive deficit diagnosed.</li></ul>



**Figure 18:** A tester sitting in front of Terabot's interface The eye tracker is located at the bottom of the screen

A pilot randomized clinical trial with random assignment to different experimental conditions was conducted. After meeting the inclusion criteria and signing a consent form, the patients participated in the study. Psychiatrists informed them about the purpose and conduct of this clinical trial and addressed any questions they had. Patients consented to have their voice and eye-tracking data recorded. It is important to note that no photograph or video of the patients could be taken during these conversations (the bioethics committee would not have approved this). The anonymity of patients participating in the study had to be guaranteed. In five sessions per week (one per day), the patients talked to Terabot. Each conversation lasted for about 7–15 minutes. The time depended on some factors, e.g. if the patient agreed to participate in the relaxation exercise. The patients could choose one of three emotions to talk about: anger, fear, or shame. The topic could then be repeated or changed the next day. A small conversation sample can be seen in Figure 19, and a bigger part of the dialogue and Terbot’s utterances are shown in Appendix A. The patients were seated in front of Terabot, with an assistant who was invited into the same room. This was required by the psychiatric hospital for the safety of the patients.



**Figure 19:** Sample of a dialogue between the patient and Terabot about the chosen emotion - here: anxiety (in Polish and English)

### 3.2.1 Eye tracker Gazepoint GP3

In parallel with testing the dialogue system, eye-tracking data were collected. For that purpose, the eye tracker Gazepoint GP3 was chosen. It collected eye-tracking data from the patients while they were interacting with the dialogue system. This stationary eye tracker was chosen

for two important reasons: firstly, this type of eye tracker does not restrict participants' movement in any way and does not require any additional items/devices to be worn on their bodies. These features are important when conducting experiments with patients with schizophrenia. As already mentioned, those patients are treated with antipsychotics, mostly in combination with other antidepressants or mood stabilizers, which may affect their behavior. As the experimental setting is in a psychiatric hospital, the clinical conditions require that the patients cannot be additionally disturbed, even in the smallest form of any wearable device.



**Figure 20:** Photo of the stationary Gazepoint GP3 eye tracker used in the study (with the biometrics toolkit also presented)

For the eye-tracking data collection (during the therapeutic sessions), it had to be guaranteed that the patients would not have to touch wires or wear any additional devices (e.g. eye-tracking glasses) during the study. Secondly, the choice of a stationary eye tracker can also be justified by the fact that the main focus of the experiments is on providing conversations. Wearing additional equipment (such as eye-tracking glasses) would interfere with the naturalness of a therapeutic conversation.

The sampling frequency chosen for the experiment was 60 Hz. Accordingly to [63], this sampling frequency can be used in the Human-Computer-Interaction domain, when analyzing fixations (described in detail in the following chapters). The selected eye tracker has dedicated software ('Gazepoint Analysis Software'). It enables the collection and analysis of eye-tracking data simultaneously with each timestamp of the audio data of the conversation. This software also enables the downloading of CSV files (which were important for the eye-tracking data analysis). Additionally, there is the possibility of analyzing videos of eye movements in each

collected timestamp. This software is user-friendly and allows fast interpretation. In Table 4, some technical specifications of the device are presented, according to the provided Gazepoint GP3 Mobile Manual. Figure 21 shows a demonstration setup with a monitor and an eye tracker at the bottom of the screen. The patients were sitting in front of the screen, where Terabot's interface was presented, as shown in the figure. The mouse was used only to turn on the operating system and the dialogue system. For conducting the conversations, it was hidden on the left side (where the assistant was sitting), being invisible to the patient.



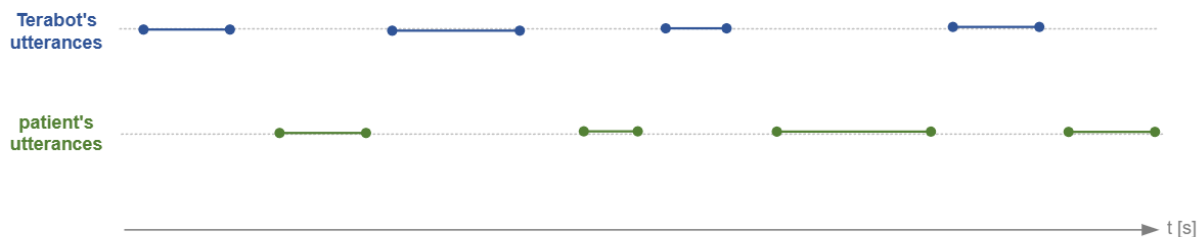
**Figure 21:** Experimental setup at the IPIN

**Table 4:** Eye tracker technical specifications, according to Gazepoint GP3 Mobile Manual (provided by the manufacturer)

Specifications	Values
Sampling Rate	60 Hz, 150 Hz
Accuracy	0.5°–1°
Spatial Resolution (RMS)	0.1
Eye-Tracking Mode	Binocular
Operating Distance	38 cm – 56 cm
Tracking Range (Head Box)	23 cm × 18 cm
Calibration	5 point
Tracking Recovery Time	< 50 ms
Size	23 cm × 28 cm × 16 cm

### 3.3 Problems identified during conversations

During the experiments at the psychiatric hospital, it was observed that patients conversed quite effectively with Terabot. The therapeutic conversations went well when patients were in a relatively stable mental state (probably due to the effects of prescribed medication). This was especially true when patients spoke at a normal speech rate with moderate loudness, and their utterances had a relatively average length (compared to healthy participants who tested the dialogue system). An ideal flow of a conversation is shown in Figure 22. After each statement the patient makes, Terabot can respond to the patient and vice versa.



**Figure 22:** Dialogue system (Terabot) response time to patient utterances - ideal situation

Some unforeseen problems occurred when conducting experiments with patients at IPIN. These issues paused the dialogue flow and sometimes caused unintended patient discomfort. The following problematic situations were identified when patients interacted with the dialogue system:

- Interruption of the patients' answers by Terabot,
- Patients having to wait a long time for Terabot's response,

- Waiting too long for the patient’s answer,
- No non-speech related information about patient participation during conversation.

These are described in detail in the Chapters 3.3.1 - 3.3.4. This dissertation addresses these issues and sets out solutions made possible by implementing an eye tracker.

### 3.3.1 Issue 1: Interrupting the patients’ answers

As it transpired, patients gave answers that were sometimes very short and sometimes very long, with many pauses during their responses. In fact, these pauses sometimes lasted several seconds. In these cases, the dialogue system did not recognize the pause appropriately, detecting it as the end of the utterance and proceeding with the dialogue flow. This is a problem which has also been pointed out in some research articles on spoken dialogue systems (e.g. [28]). In spoken dialogue systems this task is known as ‘endpointing’ or ‘endpoint detection’. It can be challenging due to background noise and the tendency of people to pause during their turn.

Consequently, Terabot’s utterances interrupted the rest of the patient’s answer. This situation is illustrated in Figure 23. It also occurred when a patient experienced a symptom known as logorrhoea [102]. Logorrhoea is a communication disorder characterized by excessive wordiness (verbosity), repetitiveness, and often disjointed, jumbled, or irrelevant speech [99, 103]. Even when Terabot posed a simple close-ended question (e.g. “Do you want to talk to me about shame?”), a patient with verbosity could not respond briefly or simply give the necessary information.

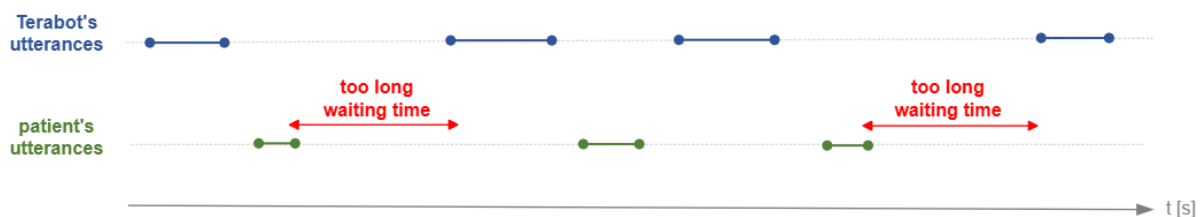
In both of these situations, the patients had not finished speaking. Consequently, their speech was cut off suddenly. From the dialogue system’s point of view, these situations prevented it from working correctly (e.g., resulting in incorrect intent recognition), and the dialogue system abruptly interrupted the patient. Based on the intentions of the already spoken utterances, the system makes a decision about an upcoming answer, instead of waiting until the end of the patient’s utterance. This is why there is a need to integrate additional information in the form of another signal input into the dialogue system, indicating that the patient’s response has come to an end.



**Figure 23:** Challenging situations for the dialogue system, during patient-Terabot conversations: interrupting the patient’s speech

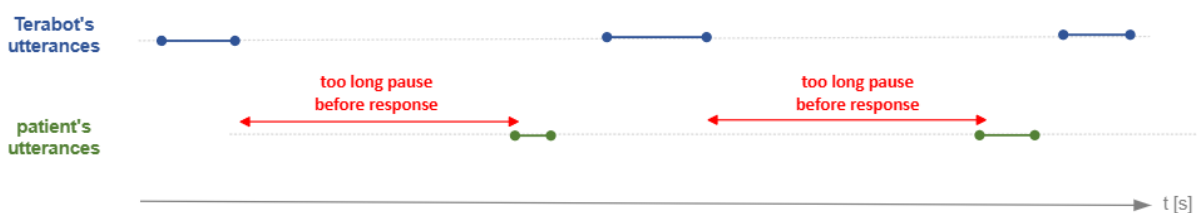
### 3.3.2 Issue 2: Patients having to wait a long time for Terabot's response

Another situation that caused the dialogue to pause is illustrated in Figure 24. The dialogue system waited for the patient's response for a very long time (the microphone was set to wait indefinitely). This occurred when patients were shy or reserved, or spoke very quietly, for example. Their utterances mainly consisted of brief phrases (e.g. "I do not remember" in Polish: "Nie pamiętam") or single-word answers (e.g. "Yes" or "No"). When these utterances were too quiet or rapid, the ASR module could not detect them as valid speech input (likely they were recognized as unintentional background noise). This caused the microphone to still be on, and the patients did not know that their answers were not recognized. It resulted in waiting for a very long time and it turned out that the patient having to repeat the answer to continue the conversation (which sometimes happened naturally after some prolonged waiting time). In terms of the flow of the whole conversation, this situation caused an unintentional pause. Such waiting made the patient feel insecure or even irritated; this is not desirable in a dialogue system designed for therapeutic purposes.



**Figure 24:** Challenging situations for the dialogue system, during patient-Terabot conversations – too long waiting time for Terabot's response

### 3.3.3 Issue 3: Waiting too long for the patient's answer



**Figure 25:** Prolonged waiting for the patient's answer

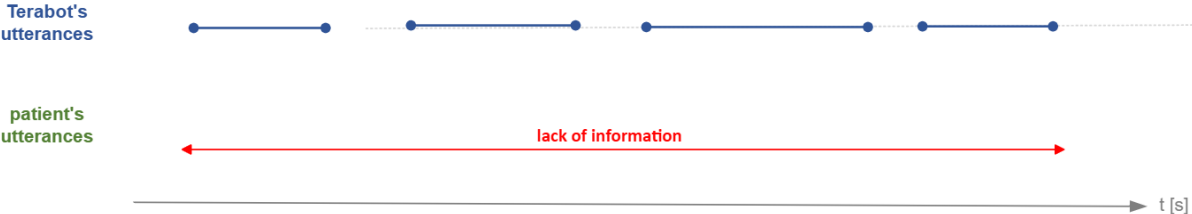
Another problem that was observed during the experiments is presented in Figure 25. There were often long periods of time where patients did not respond to Terabot during the dialogues. After a question was posed to the patient, no answer (or any speech signal) was directed to Terabot. In these situations, the microphone was still on, waiting for a speech signal to be detected. After the conversation ended, patients were asked why it took them so long to answer.

Some said they had been thinking about finding the best words for an answer. Others said they had been struggling to concentrate, their thoughts were elsewhere, and it took them a while to return to the conversation. There were also cases where, despite agreeing to participate, the patients refused to answer; they tested how the system would react to their complete silence. Again, the problem is that the patient could even leave the room without saying anything; the dialogue system would still wait for the speech signal to come. This is not the appropriate behavior of a dialogue system, so a solution had to be found to extend the functionality of this speech-only dialogue system. This dissertation aims to explore the presented issue and propose a solution in cases where implementing an eye tracker could be beneficial.

### 3.3.4 Issue 4: Lack of information about patient behavior

The experiments showed that patients can experience complicated and often overwhelming emotions when opening up to Terabot. The flow of the conversation with Terabot has been explicitly designed so that after talking about these difficult emotions, the patients can participate in a relaxation exercise. This enables them to try to calm down their emotions. The relaxation exercise lasts a few minutes and consists of therapeutic statements, during which the patient should try to focus on inner thoughts and breathe slowly, so basically try to relax. The relaxation exercise is also the time for the patients to train their mindfulness.

In the RASA-based version of the dialogue system, there was no means of contacting the dialogue agent other than speech. This means that there is no place for any feedback from the patient during the relaxation exercise. Since no speech is collected, there is no certainty that the patient is even sitting in front of the screen and listening. In terms of how patients behaved during this relaxation training, it was observed that there were different approaches. One group of patients tried to concentrate and focus on their inner thoughts, while another group was not interested in following the instructions given by Terabot. The latter distracted themselves by focusing on different things present in the room. They did not want to engage in addressing their emotions.

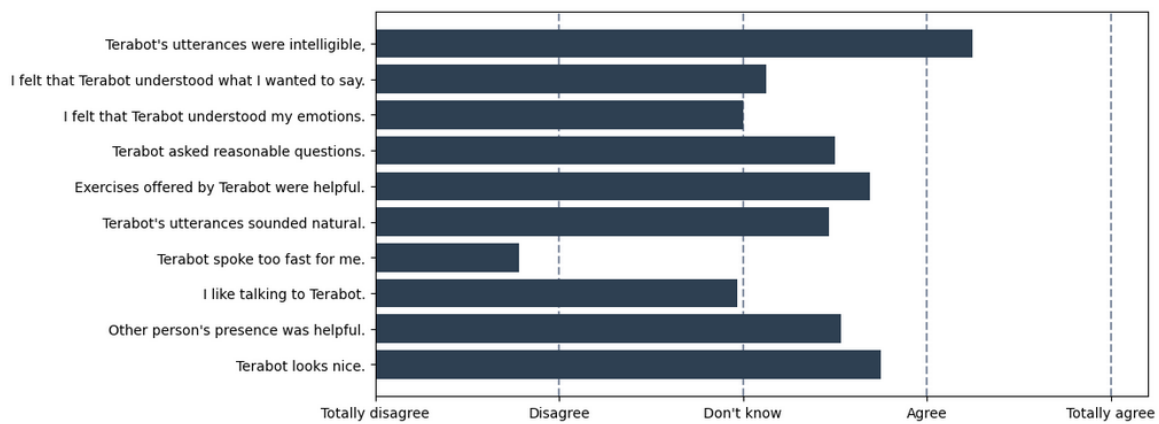


**Figure 26:** Lack of feedback on the patient’s presence for dialogue system

With the speech-only version of Terabot, an assistant needs to be physically present in the room to confirm the patient’s presence. Only then can the assistant check how the patient behaved and whether the patient has tried to follow the exercise instructions and maintained

concentration. This also applies to the whole dialogue with Terabot. This solution consumes not only money but also time. Additional staff must be hired/paid to sit next to the patient during the therapy. Therefore, it is necessary to improve the performance of the dialogue system by introducing another type of signal, which would indicate the patient's presence. It is important to gain information about whether the patients have been actively participating in the relaxation exercise or maybe even in the whole conversation.

### 3.4 Patients' satisfaction regarding conversations with Terabot



**Figure 27:** Patients' satisfaction regarding dialogues with Terabot

After the session of conversations with Terabot, each patient was asked to fill out a survey. Figure 27 shows the results of the patient questionnaires, filled in after the last therapy session. The majority of patients confirmed that Terabot's utterances were intelligible (4.2 on the Likert scale) and the relaxation exercise was found helpful (3.8 on the Likert scale). Patients also felt that their responses and emotions were understood (3.5 on the Likert scale). In general, it can be said that patients rated the quality and naturalness of Terabot's utterances highly and they were satisfied with Terabot's visual interface.

Furthermore, patients' attitudes to and expectations of Terabot varied, with some patients having a more positive attitude or higher expectations than others. Patients behaved differently in conversations: some were cautious, while others challenged the system or expected compassion. These findings were highly encouraging for our research team, as they demonstrate that dialogue systems may play a meaningful role in psychiatric therapy.

# Chapter 4

## Analyzing the need for eye tracker calibration in a spoken dialogue system

### 4.1 Problem description

As presented in Chapter 3.2.1, a stationary eye tracker was chosen for collecting patients' eye-tracking data at IPIN. This comes with a risk: the collected eye-tracking data might be affected by the free head movements of the patients. In addition, a further problem has been identified: as the study's main aim was to hold therapeutic conversations with the dialogue system, it was impossible to perform a calibration beforehand. This would have discouraged the patient from opening up during the dialogue, and their comfort would be disrupted.

In a research paper by Hareźlak et al. [104] the possibility is analyzed to shorten the calibration procedure. Other reports from this research group provide information about experiments with uncalibrated data. In these experiments, 41 people participated, and a total of 802 trials were performed [105]. Their results demonstrate that an eye tracker can be used as a pointing device for simple, well-defined tasks, without prior calibration for each user (when the task does not require point-to-point gaze mapping).

Therefore, I wanted to investigate whether the quality of the collected data without performing the calibration and without any boundaries for the patients' movements while conversing would be sufficient. This chapter presents an experiment, which was carried out to determine if the calibration process can be omitted when using the Gazepoint GP3 eye tracker for the purpose of the dialogue system.

### 4.2 Experiment methodology

The proposed experiment consists of two parts. In the first part, the diameter of the participants' pupils is measured. This is done without prior calibration of the eye tracker. In the second part, calibration is performed first, followed by measurement of the participants' pupil

diameters. In both cases, the person looks at a changing laptop background (black and white).

32 subjects participated: 16 women and 16 men (between 21 and 30 years old), who were students from the WUT and the University of Warsaw. Thirteen wore corrective lenses or glasses. During this experiment, the participants were deliberately not asked to remove them, even though wearing them might affect eye-tracking data quality. The reason for this approach was to imitate the clinical settings as much as possible. Due to the specificity of the experiments, patients in the psychiatric hospital cannot be asked to remove their lenses or glasses while conversing with Terabot, as this would have a negative impact on the therapy sessions. Patients with psychiatric disorders could feel uncomfortable during such a session, and it could also lead to a worsening of their emotional/psychological state. This is why the decision was made to allow patients to wear glasses or contact lenses, even if it introduced some problems into the collected eye-tracking data. This approach ensured that the laboratory conditions closely imitated therapy sessions with patients. Experiments were conducted at the Biometrics and Machine Learning Group laboratory at the Faculty of Electronics and Information Technology (WUT). As Figure 28 shows, the experimental set-up consisted of a laptop and the stationary Gazepoint GP3 eye tracker (with Gazepoint Control and Gazepoint Analysis software installed).

The participants were seated in an adjustable chair, and their movement was not restricted. It was important to ensure maximum freedom and comfort during the experiment because the aim was to imitate the experimental clinical settings. The distance between the participant's head and the laptop screen (and eye tracker) was approximately 40-50 cm. The eye tracker was placed directly under the screen, behind the keyboard. In this way, it was possible to detect the participants' eyes. The lighting during the experiment was natural (daylight), evenly diffused, coming in through the window, and controlled by blinds.

The task for each participant consisted of gazing at a laptop screen, during which completely black or white backgrounds were displayed in sequence. During this time, eye-tracking data were collected by the GP3 Gazepoint eye tracker. The display duration of each background was set to 5 s, except for the first black background, which was displayed for 8 s. This extra 3 s of time for the first background allowed their eyes, particularly their pupils, to adapt to the conditions, as it was the beginning of the experiment. The data from these additional 3 s were not calculated during the analysis. The total display time (for each session) lasted 28 s. Each participant took part in two sessions (with an appropriate relaxation interval between them to allow their eyes to rest for a while). The first session of data collection was performed without prior calibration of the eye tracker; while the second session was preceded by calibrating the eye tracker (meaning using the calibration function provided by the software).

For the no-calibration experiment, only the correct distance between the participant's eyes and the eye tracker was considered (using Gazepoint Control software) when the participant sat before the screen. The constant detection of the eyes and pupils in the software made it possible to determine the correct distance away from the head and eye. Feedback was given on the distance of the eyes from the device (too close/optimal/too far away), and the experiment



**Figure 28:** Experimental setup at the WUT with a participant, conducted for calibration necessity analysis



**Figure 29:** Experiment scheme: changing backgrounds for analysis of pupil size parameter; with and without calibration

only started when the feedback was positive. On the other hand, for the calibration experiment, the study was preceded by a calibration procedure provided by the Gazeport software. This procedure consisted of gazing for a few seconds at the white dots appearing on the black screen. In each corner of the screen, one dot appeared, and finally, one in the center of the screen. As already noted, this study was conducted with conditions that imitated the conditions in the psychiatric hospital as closely as possible. This meant that each participant was free to blink and move their head, and there were no restrictive rules given to the participants. In practice, it turned out that all these participants actually behaved naturally; individuals sat upright and gazed at the laptop screen without making excessive movements.

The laboratory room was quiet when conducting experiments, with no sudden noises or disturbances. All participants were in the same room for several minutes before the test, allowing their eyes to adjust to the lighting. When conducting the study on healthy participants, it was ensured that the conditions during the experiment were as constant as possible, as follows:

- **Lighting conditions:** The pupil responds to changes in light intensity [106], which is why during the experiments were conducted on days with about the same amount of sunlight, the lighting was natural (daylight), evenly diffused, coming in through the window, and controlled by blinds.
- **Cognitive load:** as it is described in literature, cognitive load and concentration levels [107, 108] have an impact on pupil behavior. This is why participants took part in the study on the same day, after attending in the university's laboratory. It was a class in the middle of the day. After a few minutes of relaxation and preparation, participants were asked if they were ready to take part in the experiment.
- **Health status:** a brief survey was conducted to assess self-esteem and health status, asking "How do you feel?". The responses were positive/neutral ("well" or "ok"). Some people did not confirm that they were feeling good, so they were excluded from the study.
- **Alcohol/substance abuse:** consumption of alcohol and drugs affects the pupil size [109]. The effect of alcohol, stimulants, drugs, etc., was most likely eliminated since the experiment took place after the laboratory class. Participants were observed during the class and showed no signs of abuse of alcohol or psychoactive substances.
- **Emotions:** it is described that perceived emotions (also caused by sounds) have an influence on the pupil size [110, 111]. A calm atmosphere was maintained in the laboratory, and a short, quiet period was imposed on all participants just before the test. This ensured a standardized condition and minimized the potential influence of emotions on the test results.

### 4.3 Results and discussion

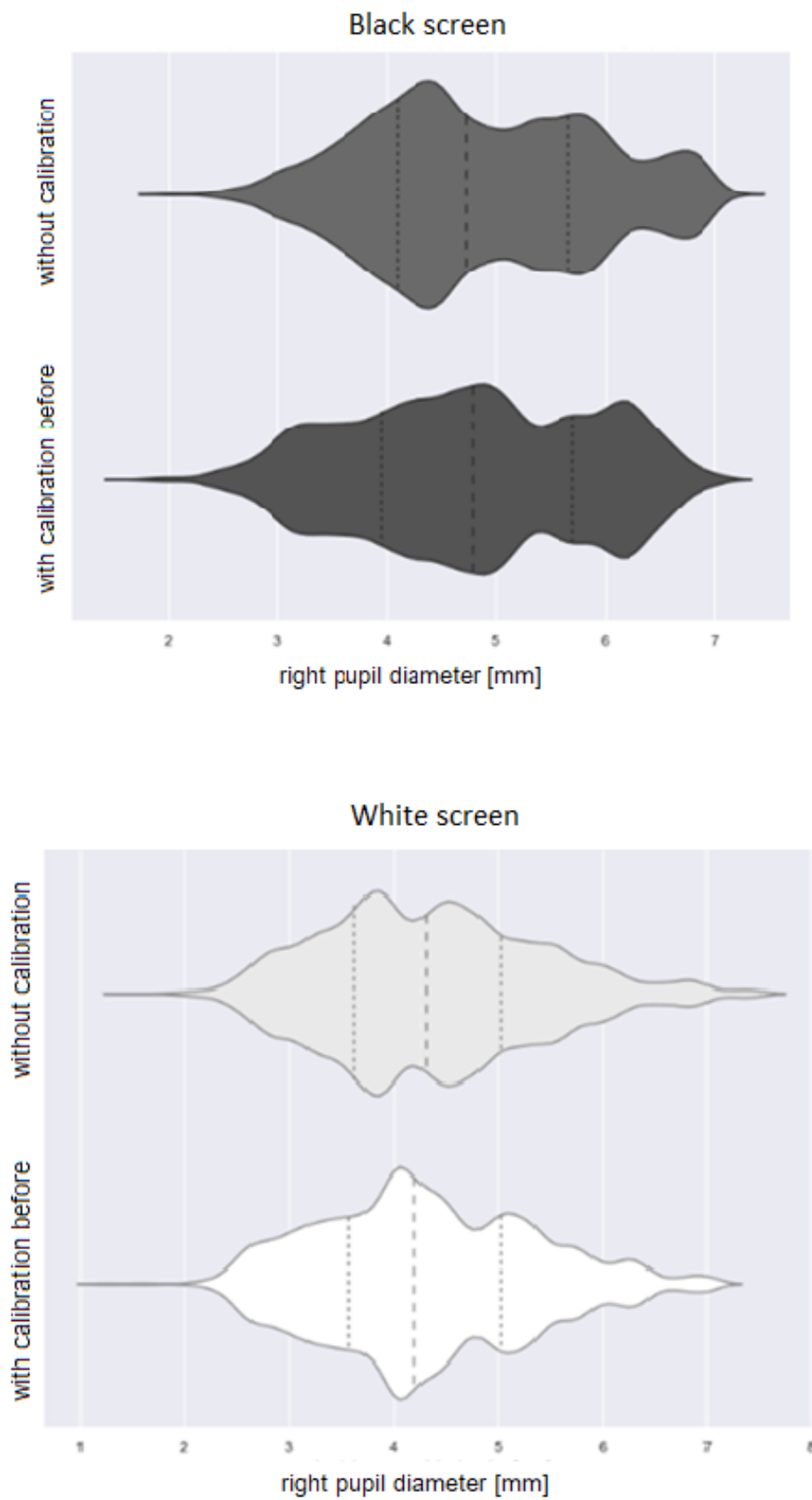
The resulting eye-tracking data were analyzed after experiments were carried out with and without calibration. According to other studies in the eye tracking domain (e.g. [112]) when analyzing data concerning pupil sizes, it is advisable to skip the first few seconds/minutes of the experiment (depending on the duration of the entire study). This is because it takes a few seconds for the eye to adapt to the conditions (such as adaptation to display brightness, novelty effects, etc.) when it first looks at the screen.

In the case of our experiments, the first 3 s of the first black background in the experiment were removed for each subject. As it is known from the research articles about pupillometry (e.g. [113]), there is a mechanism known as the consensual pupillary light reflex. It is the constriction of the pupil in the other eye when light is shone into one of them. This is because signals from the illuminated eye travel to the brain, triggering constriction of both pupils. Due to this mechanism, it can be assumed that the other pupil functions in the same way, unless stated otherwise (e.g., due to eye diseases). This is the reason the data analysis focused only on the right pupil diameter size (of each participant) for each background separately.

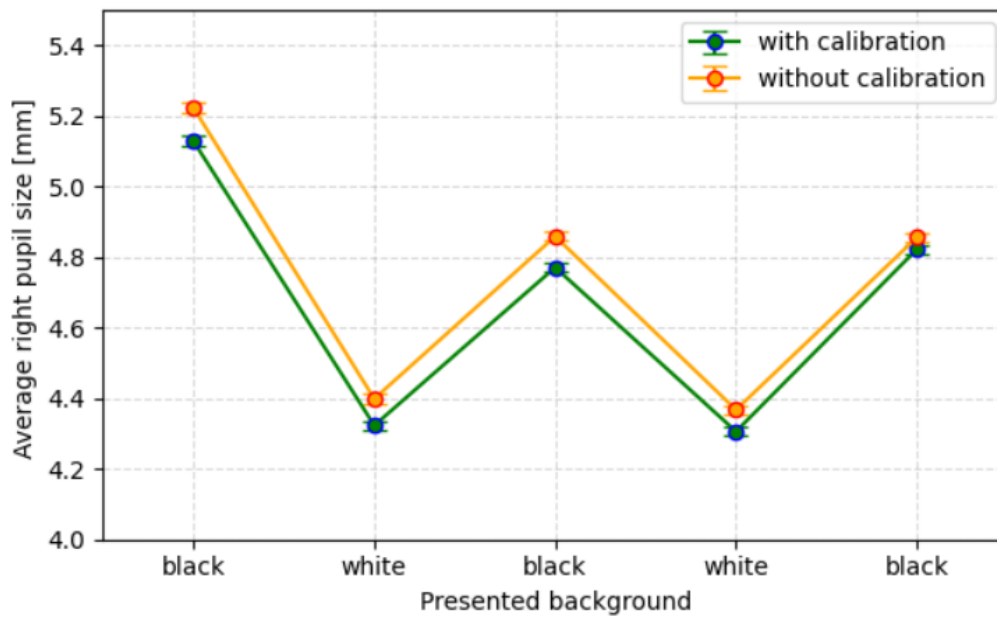
Figure 30 shows violin plots for the two selected background types: one black background and the following white background. As can be seen, the average pupil diameter of all the participants has decreased with the presentation of a white background (for the black background, it was about 4.8 mm; for a white background, it is about 4.1 mm). The protective function of the eye explains this change of pupil size: when the light coming into the center of the pupil changes to a much brighter light, the pupil constricts to protect the inside of the eye. For each set of graphs presented in Figure 30, it can be seen that the pupil diameter data have (almost) identical ranges, as well as median, first, and third quartile values (analyzing each background separately).

Figure 31 shows the average diameter of the right pupil of all participants for each type of background and each experiment (with and without calibration), along with the corresponding calculated confidence intervals. In this case, the confidence level was set at 95%, as usually recommended [114]. As can be seen, the average pupil size values and their corresponding confidence values indicate no difference in the data derived from the experiments with and without calibration. There is only a difference in the calculated averages between the black and the white backgrounds (whichever is used), showing that the participants' eyes accommodated to the background present on the screen.

It can be seen that the pupil size values from the experiments with pre-calibration are minimally (negligibly) lower: the differences between the obtained data are approximately 0.1 mm at most. This is not critical because the differences for each background can be seen very clearly in the data with and without calibration. This is an important point as it makes the hypothesis that no calibration is required very plausible.



**Figure 30:** Violin plots for right pupil diameter data, in the upper figure for one black background and in the lower figure for one white background



**Figure 31:** Average values of participants’ right pupil diameters with their corresponding confidence intervals for each background presented separately

## 4.4 Conclusion

As shown in Figure 30 and Figure 31, the graphs of the average data collected concerning pupil diameter show a clear trend that supports the hypothesis that calibration is not required. It indicates no significant difference between the data collected from the experiments with and without pre-calibration. Obviously, validating this hypothesis would require a larger study and confirmation through statistical tests. However, the results strongly suggest a high probability that calibration of this eye tracker is not needed. In the study, the focus was on analyzing pupil diameter and the necessity of calibration. If comparing pupil data, where size and variation are measured in hundredths of a millimeter, there was no significant effect from calibration. It is supposed that other parameters are likely to retain the same level of information regardless. The experiment confirmed that calibrating the stationary Gazepoint GP3 eye tracker is not a critical requirement. This means that the eye tracker can be used with a dialogue system without prior calibration. **The research objective “[RO1]: Investigating whether an eye tracker needs to be calibrated when used in a spoken dialogue system” has been successfully achieved.** The results of this chapter have been presented in articles [115, 116].

# Chapter 5

## Analysis of the gaze of a patient conversing with Terabot

### 5.1 Problem description

As described and explained in Chapter 3.3.3, two problematic situations (related to Terabot giving responses) were detected during the clinical study. Briefly, the first one occurred when the patient's speech was too quiet or too fast for the ASR module to recognize. As a result, the dialogue agent did not react: the microphone was on and waiting for a speech signal to be detected. The patients faced a prolonged wait and had to repeat their response. The second issue occurred when patients spoke for too long (e.g., because of logorrhoea) or when there were many pauses of several seconds in a patient's utterance. In this scenario, it was often the case that the dialogue system interrupted the patient while they were answering. The dialogue system decided on the response topic (for Terabot's utterance) based on the intentions of the already spoken utterances. These two situations challenged the dialogue system, as the patient had not yet completed the utterance. These situations may have disrupted the entire therapy session and impaired the dialogue system's ability to appropriately respond to the patient's behavior. What was needed was an indicator (other than speech) that could provide information about the finished utterances.

### 5.2 Methodology

To analyze their gaze behavior while patients ended their utterances, 72 utterances of 9 patients with corresponding eye-tracking data were used. Regarding the analysis of the speech signal and the corresponding eye-tracking data in each timestamp, it should be noted that manual synchronization was not required, as it was already done by the provided software ("Gaze-point Analysis") in the settings of the eye tracker. Each time the eye tracker was turned on, it collected not only eye-tracking data but also audio data (at the same time). This means that for

each timestamp in each conversation, the collected eye-tracking data (e.g. fixations, pupil diameter) could be associated with the corresponding timestamp in the output files (video and audio recording). This allowed for precise information (about where patients gazed while conversing with Terabot) to be collected. As mentioned in previous chapters, obtaining good-quality eye-tracking data from all patients who participated in this study was difficult. The reason was a lack of concentration in head and body movements. Despite this limitation, it was still possible to analyze the collected data for a subset of the patients.

During conversations, it was analyzed whether the parameter Fixation Point of Gaze (FPOG) provided non-verbal information about the end of patients' utterances. The parameter FPOG is composed of two coordinates (FPOGX for X and FPOGY for Y) of each user's fixation point on the screen (at each timestamp). An analysis of where these fixations occurred most frequently when patients reached the end of their utterances had to be conducted. The hypothesis was that if a specific region of Terabot's interface is focused upon (while patients finish speaking), it could help predict where the fixations would be when the patients come to the end of their utterance. As is known from human communication research [117, 118], whenever people are talking, the person who is speaking starts to look more at the interlocutor as the end of the utterance approaches. This is because the speaker thinks about what to say and how to say it (the concentration is on the thoughts and speech). When the speaker is near the end of the speech, looking at the interlocutor is a nonverbal signal that he/she is about to stop talking, pay attention to the other person, and now wait for a reaction. This approach was applied to the dialogue system, making the conversation with the Terabot more human-like. Although previous studies on turn-taking behavior have focused on mentally healthy populations, I wanted to explore if and how these findings could be applied to individuals with psychiatric disorders.

### **5.3 Utterance and eye-tracking data analysis**

The first step consisted of analyzing patients' utterance data. The endings of each patient's utterance had to be found, which will be further described as closing phrases of the last spoken sentence. These closing phrases were manually marked individually for each patient's last spoken sentence. Usually, they consisted of the last three words of the last spoken sentence. Figure 32 shows graphically the closing phrases of a patient's last spoken sentence in the utterance.

The next step considered the fixation data. Terabot's visual interface had to be divided into areas of interest (AOIs) to get information about fixations in these selected regions. The Gazepoint GP3 software used in the study collects eye-tracking data and allows users to specify which AOIs in the presented media are of particular interest. It is possible to manually define these AOIs by specifying the fixed corners of each one, which are the same for the whole recorded experiment: fixed coordinates of the displayed media. Depending on the purpose of the study, the number of AOIs and their size, the coordinates can be set individually for the

Pacjent: Tak, przypominam sobie (...).  
Poczułem wtedy niepewność, byłem nieswój,  
obawę, niepokój, to mi towarzyszyło głównie.

Closing phrase

**Figure 32:** Example of a patient's utterance and the closing phrase

entire recorded experiment. The results are fixation data which are grouped by the specified AOIs; the data consists of the coordinates of each fixation point in those regions, the fixation duration, and the order in which they occurred.



**Figure 33:** Photo of Terabot's interface with selected Areas of Interest (AOIs), such as green AOI – person

Figure 33 shows Terabot's interface with eight selected specific AOIs, including not only Terabot's person but also other subjects present in this visual representation (bookshelf, plant, etc.). Considering that the experiment involved patients with psychiatric disorders, a question arose as to whether any objects could contribute to distraction during the patients' speech. Here is a list of the selected AOIs for analysis:

- 'AOI writing' – the writing at the top, which is introduced to start the conversation,
- 'AOI person' – the person representation,
- 'AOI buttons' – the button and microphone icon at the bottom,

- ‘AOI plant’ and ‘AOI bookshelf left’ – subjects related to Terabot’s environment on the left,
- ‘AOI bookshelf right’, ‘AOI vase’, ‘AOI books’ – subjects related to Terabot’s environment on the right.

Studies in psychology (e.g. presented in articles [119, 120]) have shown that generally the eyes are one of the first and most frequently fixed regions within the human face. As Terabot is a dialogue system for patients with specific psychiatric disorders, the result of the fixation analysis is important – the aim was to adapt Terabot’s interface to the patients’ needs as much as possible (not to healthy users). This is why it could not automatically be assumed that the patients’ gaze would be mostly fixated on Terabot’s eyes or even face.

In therapy sessions, and even when considering everyday conversations, when people discuss difficult topics with much emotional load, for example, those related to shame or anger, even healthy individuals do not always maintain direct eye contact with the person they speak to; they might simply look away. This could be caused by the intense emotions involved in such conversations. This is why the decision was made not to choose Terabot’s eyes as a small, separate AOI. There are also other reasons that show why the eyes have not been decided to be of special interest when analyzing patients’ data:

- **psychiatric illness**, which might affect the patients perception and concentration level;
- **medication** the patients received during their treatment and also during conversations with Terabot, which might have an impact on gaze behavior;
- **therapeutic session environment**, which is quite private and delicate.

In the case of therapeutic sessions, it is important to consider whether patients were gazing at Terabot (‘AOI person’) or maybe not this region at all (all other AOIs).

**Table 5:** Fixation durations in AOIs of Terabot’s interface during closing phrases.

Area of interest (AOI)	No. of fixations	average fixation duration [s]	max. fixation duration [s]
<b>AOI person</b>	<b>2821</b>	0.54	<b>2.0</b>
AOI buttons	347	0.42	1.8
AOI bookshelf left	564	0.37	1.13
AOI writing	309	0.34	0.9
other AOIs	209	0.21	0.53

Reports concerning the human communication literature [118] indicate that after people finish speaking, they still look at the interlocutor while waiting for the response. This is why I included the first 3 s immediately after patients finished speaking in the analyses. This value

was adjusted experimentally and it was decided that 3 s was a reasonable time to wait. Table 5 shows the collected fixations while the patients finished speaking (in the different AOIs of Terabot's interface) with these additional 3 s included.

## 5.4 Results



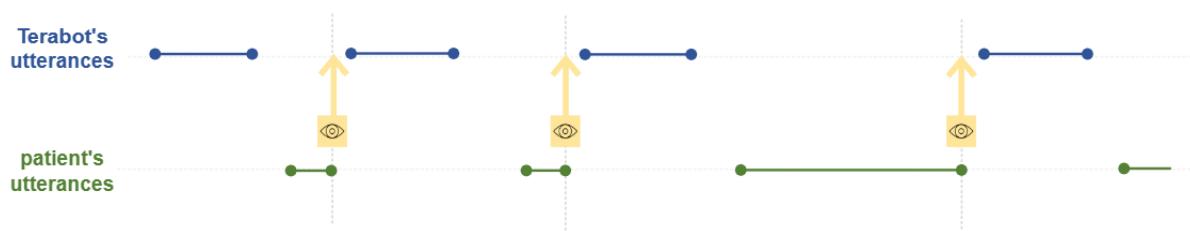
**Figure 34:** Visualization of Terabot with color intensity indicating viewing frequency: stronger colors – longest fixation numbers and duration; faded colors – fewer fixations and durations

After analysis, the area with the highest number of fixations was ‘AOI person’ (with 2821 fixations), with an average fixation duration of 0.54 s and a maximum fixation duration of 2.0 s. All other AOIs were investigated and the results are shown in Table 5.

A closer look should be taken at the ‘AOI buttons’ area regarding average fixation times. These symbols have no meaning from a conversational point of view, but here the average and maximum fixation durations (0.42 s and 1.8 s, respectively) are comparatively high. The reason for this might be the changing color of the microphone symbol during the conversation. It was black when Terabot was talking (indicating that the microphone was off, so Terabot was talking and not listening), and it turned red when the microphone was on to record the patient's speech. The fixation data indicate that the changing color might have accidentally caught their gaze. This information is important for Terabot's further development regarding the dialogue system's interface: possibly adjusting the colours of these buttons so as not to attract unnecessary attention from the patients. The analysis indicates the importance of the “AOI person” for the dialogue system, where most of the fixations occurred during patients’

closing phrases.

Figure 34 uses color intensity to show the AOI which the patients looked at most during their closing phrases. This intensity reflects the distribution of visual attention, with more saturated regions representing more frequently observed areas, and faded regions indicating areas viewed less often. When a speech signal (patient's utterance) is detected and the closing phrase is spoken, fixations on this AOI may provide a reliable indication that the patient has finished speaking. Figure 35 illustrates how the fixation signal can be applied to indicate the end of patients' utterances when the speech signal ends. The results of this chapter have been presented in the research articles [121,122]. It can be stated that **the research objective “[RO2:] Analysis of areas of interest in the graphical interface of the dialogue system” has been achieved.**



**Figure 35:** Schematic representation of the use of the Fixation Point of Gaze (FPOG) parameter as a possible response activation indicator for the dialogue system

## 5.5 Clinical use of the proposed solution

Based on the results achieved, I proposed to use eye-tracking data to activate each Terabot utterance when the patients' speech ended, and at the same time, their gaze was present in the specified AOI, as shown schematically in Figure 35. This mechanism of gaze-activated responses seems very helpful for indicating the closing phrases; however, there might be a considerable risk. This solution, although supported by the results of data analysis presented in the previous section, could potentially make Terabot less accessible to patients with schizophrenia. In therapeutic practice, it is not surprising that patients do not maintain eye contact with the therapist throughout the conversation. This might happen when patients experience high levels of anxiety (including social anxiety), or they feel embarrassed or distrustful (e.g., with increased delusions). For psychotherapists, eye contact is considered an important feature from a therapeutic point of view. In everyday practice, when patients do not look at their therapists, the conversation is continued by them. From clinical practice<sup>1</sup>, at the start of a therapeutic session, the therapist should proceed with the dialogue despite the lack of the patient's eye contact. However, as therapy progresses, the therapist should gently bring the issue to the patient's attention by asking:

<sup>1</sup>The co-working psychiatrist and psychotherapist, Izabela Stefaniak, PhD, MD, shared this information for this study and it can also be found in other online sources for psychotherapists, for example [123].

- “I have noticed that maintaining eye contact seems challenging for you. Could you share what might be causing this?”
- “How do you understand your difficulty with eye contact?”
- “Have you experienced this for a long time?”
- “Has anyone ever pointed this out to you before?”
- “How can I support you with this?”

This is why the decision has been made not to implement this solution to enhance Terabot’s functionality. Eye contact plays an important role in therapy, serving as an additional communication channel with the patient, but it cannot be considered a mandatory condition for psychotherapy sessions. Therefore, another solution had to be investigated. Despite that, **the research objective, “[RO2:] Analysis of areas of interest in the graphical interface of the dialogue system” was achieved.** As can be seen, the severe psychiatric disorders of the patients are a challenge not only for the psychiatrists but also for researchers and engineers who want to support them with IT solutions.

# Chapter 6

## Designing a gaze-dependent, LLM-based dialogue system for therapeutic purposes

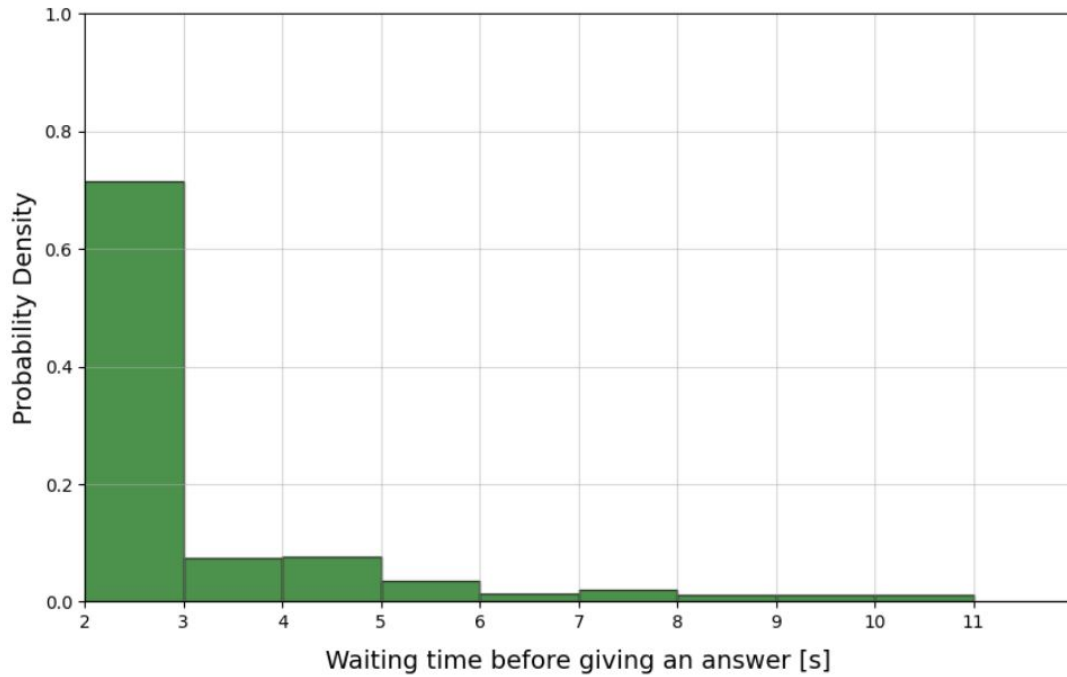
### 6.1 Problem description

As stated in Chapter 3.3.3, when conducting the experiments at the psychiatric hospital, some unforeseen problems occurred during the conversations between the patients and Terabot. After Terabot asked a question, some patients did not respond for a long time. Meanwhile, the state of the dialogue system was the same - the microphone was on, and the system waited all the time for a speech signal to be detected. In those cases, the natural flow of the therapeutic dialogue between the patient and Terabot was paused.

### 6.2 Analysis of patients' waiting time

The dialogues between Terabot and patients were analyzed, especially the waiting times between the end of Terabot's question and the start of the patients' reply. It was calculated that the average prolonged waiting time for the patient's response lasted 6.3 s. It is important to note that the standard patient waiting times of 0-2 s (before replying) were not included in this calculation. The longest waiting times lasted between 10 s and 32 s. Figure 36 shows probability density bars, which indicate how long the patient waited before giving an answer.

All in all, there were 108 cases where the patients waited longer than 2 s before replying to Terabot's posed question. This means that the conversation flow was paused in all these cases, since no speech signal had been detected. The system was waiting for the answer for the entire time, with the microphone on. There was no information or indication that the patient was still in the room (besides the information given by the assistant sitting beside the patient). These situations clearly demonstrate that speech may not be a sufficient input signal into the dialogue system. Another input signal into the spoken dialogue system should be considered to indicate the patient's behavior during conversations. I propose that adding the eye-tracking signal is a



**Figure 36:** Probability density bars of waiting times

solution to address this issue.

## 6.3 Using real-time eye-tracking data to improve the dialogue system

To solve the described problem, I propose implementing a dialogue system that uses real-time eye tracking data to indicate how patients behave during conversations. This has been identified as the next step for Terabot. Currently, LLMs are increasingly being used for dialogue system management (as previously described in Chapter 2.2.3 and so the new version of Terabot will also be LLM-based.

### 6.3.1 Using OpenAI's Assistant API for dialogue management

The GPT-4o-mini model was used as the core for dialogue management. The OpenAI's guidelines for creating a customized assistant were followed [124]. The specifications for the desired behavior of Terabot were written in the system instructions, where the detailed conversation flow and the expected tone of Terabot's answers were described. Here is a brief summarization of some of the instructions:

- Step 1: Begin the conversation with the user with a friendly greeting and ask if the user gives you permission to talk about the emotions with you.

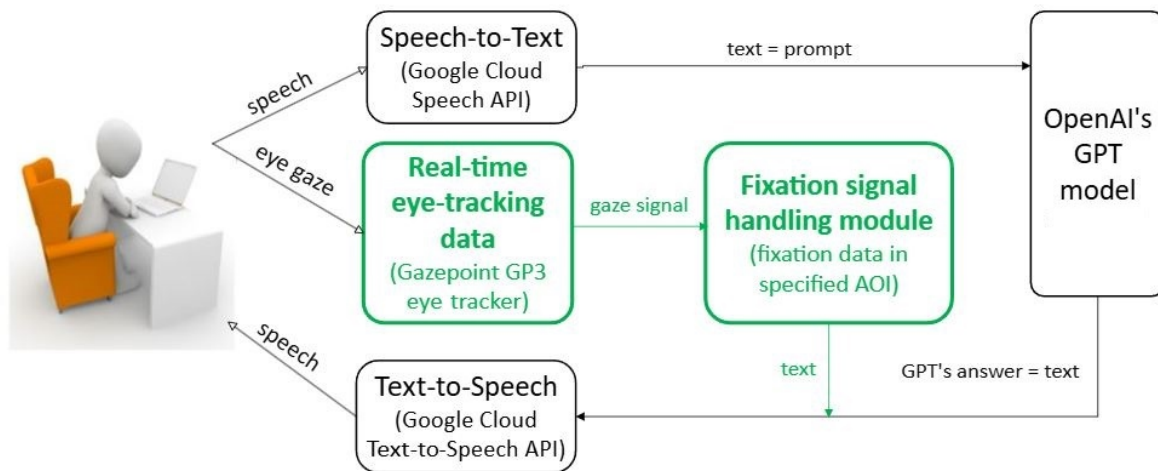
- Step 2: If the user agrees, ask which emotion should be discussed. The user can choose from anxiety, shame, or fear.
- Step 3: Once the emotion has been chosen, you can start asking questions about the chosen feeling. Ask questions like a psychotherapist. Encourage the user to share their thoughts and experiences, but be gentle and avoid pressuring them too much. You can ask 3 to 5 questions about the situation described by the user when the feeling occurred; always ask the following questions:
  - Do you remember a situation in which they experienced this emotion? (Ask the user to describe this situation.)
  - How did you feel in this situation?
  - What were your thoughts in this situation?
- Step 4: Encourage the participant to take part in a relaxing exercise. Ask if the patient would like to do a relaxation exercise together.

The model configuration parameters for the LLM behavior (meaning *temperature* and *top\_p*) were experimentally chosen to be suitable for the therapeutic purpose of conversations. *Temperature* is described as a parameter responsible for “randomness” that is used by the LLM when it generates text. When the temperature’s lower, the text output is more predictable and consistent. When the temperature’s higher, the LLM introduces more freedom and creativity in the generated output [125, 126]. The parameter *top\_p* is responsible for the control of how many possible words the model considers during text generation. Higher values of *top\_p* allow the model to include less likely options, increasing the diversity of the output [125, 126].

As stated in [125], combining temperature and *top\_p* can produce a wide variety of text styles by the LLM. When using low temperature and high *top\_p* parameters, the LLM can produce output texts which are logical and consistent due to the low temperature, but can still be rich in vocabulary and ideas because of the high *top\_p*. This combination is suggested for clear, interesting educational or informative texts. After careful consideration and experiments on the LLMs’ output texts, the most appropriate values in the case of Terabot are: for temperature = 0.15 and for *top\_p* = 0.4.

For the purpose of generating more adequate answers for therapeutic purposes, the Retrieval Augmented Generation (RAG) technique was also used. The RAG enables improvement of the performance of an AI model by connecting it to external knowledge bases and adapting generative AI models for domain-specific tasks. This helps to fill gaps in a model’s knowledge base, enabling it to provide more accurate and relevant answers [127]. The text files that describe the behavior and dialogue flow of the older Terabot version were used as RAG files in the newer dialogue system (e.g. NLU, stories, and rules).

### 6.3.2 Spoken dialogue system with eye-tracking data input



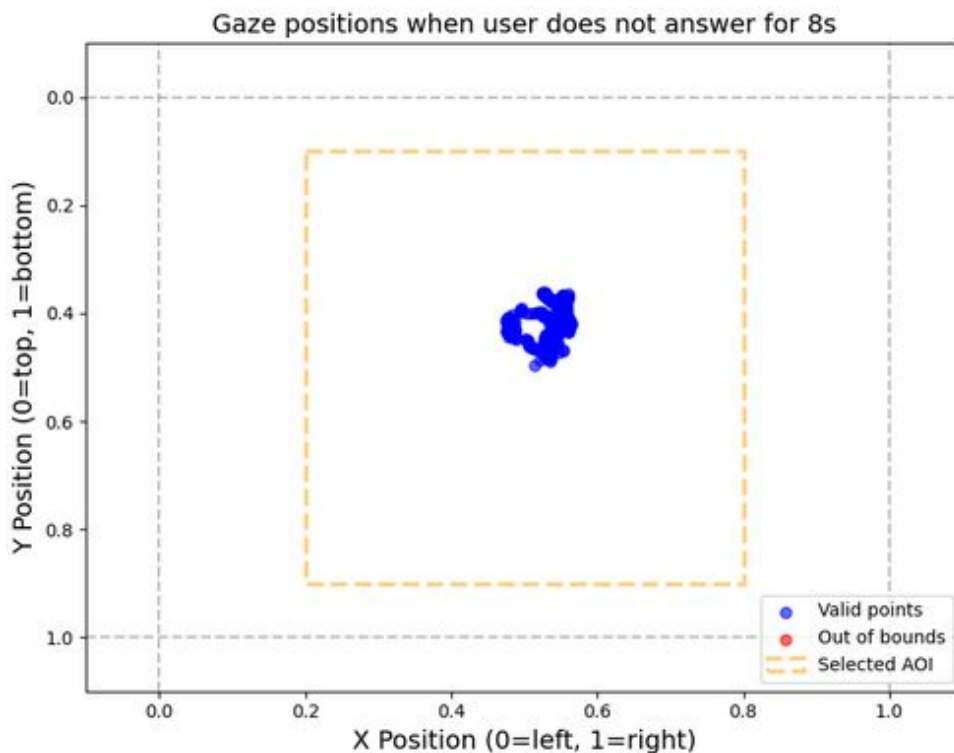
**Figure 37:** Schematic representation of the new, multimodal dialogue system: with the core of the LLM (OpenAI's GPT-model) and real-time eye-tracking data

As already mentioned, real-time eye-tracking data was integrated as additional input into the dialogue system. The architecture of the new, multimodal dialogue agent is shown in Figure 37. As the first step, when a patient talks, the speech is converted into text by the Speech-to-Text module, using the Google Web Speech API with the Polish language. After that, this text is then given to OpenAI's GPT model as an input prompt (with OpenAI's Assistant API specifically adapted for Terabot's behavior). At the same time as the speech signal is received and processed, the eye tracker gives information in the form of patients' real-time fixation data (position and duration) to the dialogue system. The fixation-signal-handling module simultaneously processes this data stream, illustrated in the right green rectangle of Figure 37. It checks if the received gaze data is within the bounds of a selected area of interest (Terabot's person area) for a specified time. It is important to note that, as this is a dialogue system, speech is considered the main signal. This means that the eye tracker serves as an enhancement of the dialogue system. It is designed to deal with situations with no speech signal for the dialogue system to process (meaning the patient, for some reason, does not speak for a long time). The idea is that the fixation-signal-handling module will only send another text to the patient (in the form of Terabot's utterance) if no speech is detected for a specified amount of time and in the selected area (experimentally determined, see Chapter 5). On the other hand, if the patient's speech is detected, meaning the patient responds to the questions posed, this module will not be activated. The text from the fixation-signal-handling module or the answer from the GPT model is then transformed into a speech signal using the Google Cloud Text-to-Speech service. The same service as in the old version of Terabot was used, Google Cloud Text-to-Speech API.

## 6.4 Offline tests – based on patients’ data

To simulate the patient’s gaze behavior as closely as possible, the gaze data from the conversations collected during the study (see Chapter 3.2) was used to test the performance of the new multimodal system.

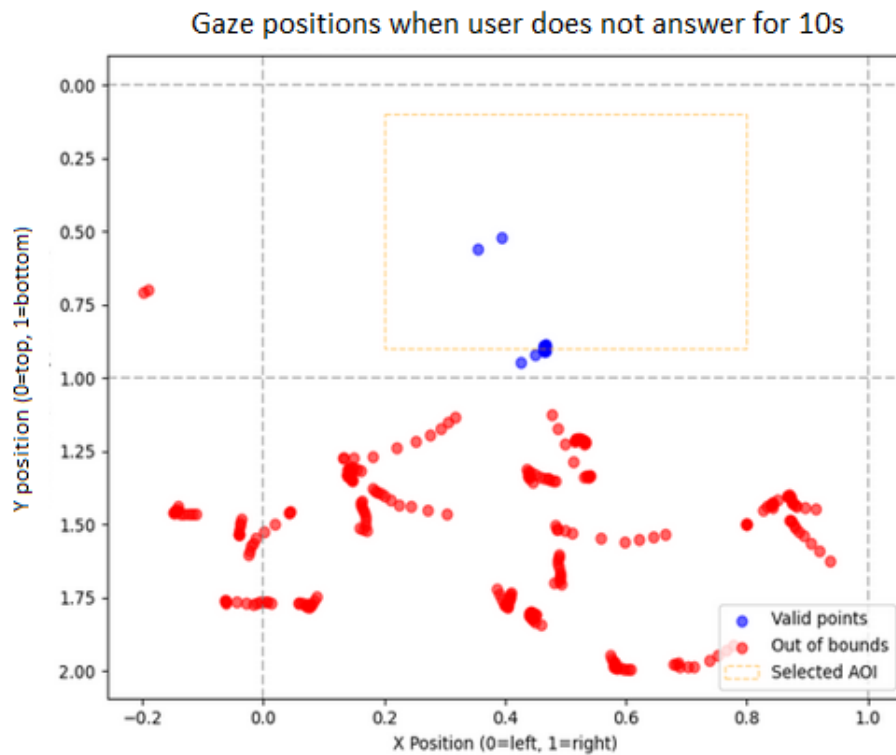
Since the new gaze-enhanced, LLM-based system responds to users’ gaze patterns, I analyzed its performance using patient fixation data. To do this, fixation points and their continuous durations were examined when the patients did not respond. Figures 38, 39, and 40 show three different, typical gaze behavior scenarios when the patients did not respond. In each of the figures, the orange rectangle indicates the area of Terabot’s person (described in the figures as AOI), whereas the grey line indicates the area of the screen.



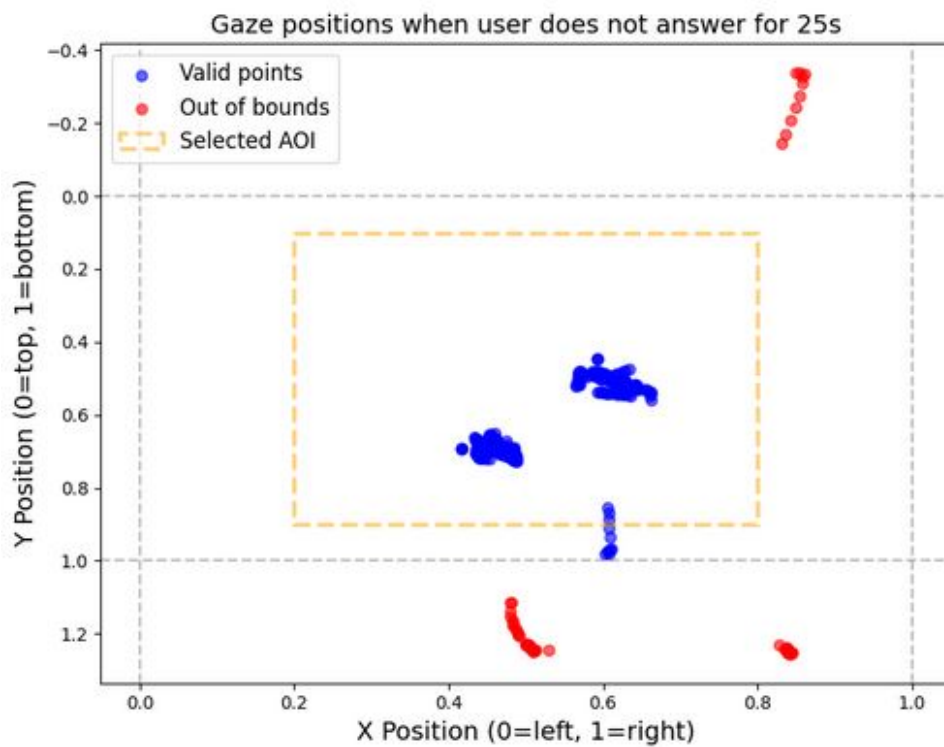
**Figure 38:** First type of gaze behavior cases (case A) during the prolonged waiting times for answers: all gazes within Terabot’s person area

The first gaze behavior case (Case A), shown in Figure 38 shows the situation when all fixations occur at the Terabot’s person. Since the patients’ gaze was present in this specific case for 8 s (longer than the selected threshold), there is no doubt that the encouraging (E) utterance would have been activated.

Another gaze behavior case (Case B) is presented in Figure 39, where the majority of fixations occurred outside of the screen (and additionally also some of them on Terabot’s person). In this scenario, when the sum of all continuous fixations outside of the AOI would last longer than the selected threshold, the presence check (PC) utterance would be activated. This logic is presented in Figure 41.

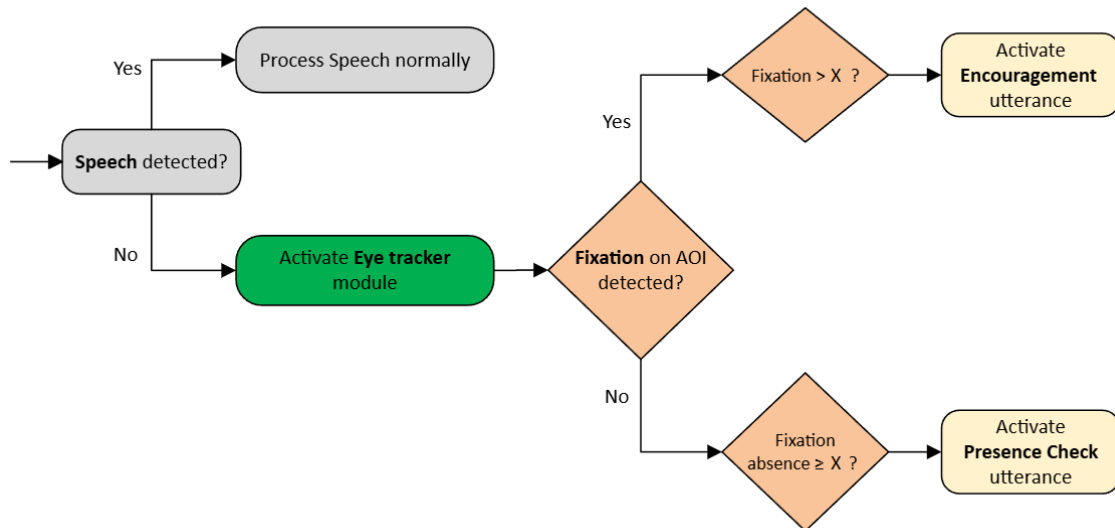


**Figure 39:** Second gaze behavior case (case B) during the prolonged waiting times: gaze almost always outside Terabot’s person area.



**Figure 40:** Third gaze behavior case (case C) during the prolonged waiting times: gaze both within Terabot’s person area and outside the screen

In the third gaze behavior case (Case C), shown in Figure 40, the fixation fluctuates between the area of Terabot’s person, the screen, and outside of the screen. In this case, it is also important to examine the sum of all continuous fixation durations closely. The E utterance or PC utterance would be activated if the gaze in the specified area lasts longer than the proposed threshold.



**Figure 41:** Gaze-response decision logic, where X – waiting threshold [s]

Table 6 shows the analysis of all the cases (with the prolonged waiting time) and the proposed threshold values for the utterance activations. When the proposed threshold is set to 7 s for E activation, only twice would the utterance have been activated. With the same threshold value, the PC utterance would have been activated eight times. This means that, all in all (when choosing 7 s as the threshold value), in 10/28 cases, there would have been an activation of the eye-tracking related utterances; that is approximately 36% of all prolonged waiting time situations.

The proposed threshold can be set at 6 s to check if more successful utterance activations would happen. When considering 6 s as the threshold value, it turns out that more successful activations (for encouragement and presence checking) would take place. The number of E activations would rise to 4, and the number of PC activations would increase to 14. This means that in 18/28 situations, where the patient does not answer for a prolonged time, an eye tracker would help. Thus, in approximately 64% of all these situations, the dialogue flow could be helped using the eye tracker.

It may be worth considering whether the threshold should be lowered further. A test with a 5 s threshold has been performed, and it has been found that the number of successful utterance activations did not increase compared to the 6 s threshold.

Therefore, I suggest that since this is a dialogue system for therapeutic purposes, this threshold should not be lowered further. The reason is that every patient should have time to con-

**Table 6:** Waiting threshold values and the belonging numbers of Terabot’s utterance activations.

Waiting threshold	No. of successful encouraging activation	No. of successful checking activation	% of all prolonged waiting time situations, with successful activations
7 s	2/28	8/28	36%
6 s	4/28	14/28	64%
5 s	4/28	14/28	64%

concentrate and focus on the inner thoughts. A waiting time of just 3, 4 or 5 s could prevent the creation of a calming and friendly environment, which is important for therapy sessions. Instead, it would be perceived as pressure to respond, with no time for reflection, which is inconsistent with the character of therapeutic conversations. As the analysis in Table 6 shows, the optimum threshold value should be set to 6 s for the dialogue system, according to the offline tests based on patients’ gaze data. This proposed value can also be further justified because the average prolonged waiting time was calculated as 6.3 s (see Chapter 6.2). As a result of the improved version of the dialogue agent using the eye tracker, this solution, with the suggested threshold value, would allow most of the dialogues to return to their usual flow.

## 6.5 Real-time tests

As described in the previous chapter (Chapter 6.4), offline tests (based on patients’ gaze data) succeeded when applied to the multimodal dialogue system; however the most desired approach for dialogue system testing is verifying its performance with humans in real-life scenarios. To compare the multimodal dialogue system with the previous version, it would require conducting a whole new study on patients with schizophrenia. For now, performing such tests is impossible (they are costly, time-consuming, and require adequate resources). Additionally, for ethical reasons, this would require a new Ethics Committee’s approval; therefore, another approach to testing was chosen. It is based on medical teaching research [128].

### 6.5.1 Participants: simulated patients and “Teatr PW” theater actors

In the medical domain, for teaching and assessment purposes and to standardize clinical examinations, “simulated patients” (SP) are used. Sometimes, the term “standardized patients” can also be found. People with this training can realistically portray patients with specific conditions (sometimes in a standardized way where they give consistent presentations) [129]. In the Polish academic community, this has been carried out at a medical university. This practice

is observed in the examination of medical students, especially future psychiatrists. Actors and actresses play roles of patients with defined medical and psychiatric disorders. During an exam, the medical student aims to provide a proper diagnosis based on the patient's behavior, speech style, etc. There is scientific evidence that SPs can be a valuable addition to psychotherapy education for psychiatry residents. It is not only considered a practical teaching and assessment method, SPs can also provide valuable learning opportunities without raising concerns associated with recording actual patients [128]. Therefore, the decision was made to test the multimodal dialogue system with the help of actors and actresses of the WUT's Theater ("Teatr PW" theater) [130].

The "Teatr PW" theater consists of WUT students who spend their free time learning to act professionally. The Director of "Teatr PW" theater is Grzegorz Sierzputowski, a Polish theater and film actor, teatrologist, director, and cultural animator [130]. The actors were asked to play the role of patients with psychiatric disorders.

### 6.5.2 Experimental setup and methodology

The tests took place in a closed, quiet room, with the experimental setup looking very similar to the setup provided at IPIN (see Figure 21 in Chapter 3.2). The two photographs in Figure 42 show actors of the "Teatr PW" theater conversing with the dialogue system. The participants sat in front of a screen with Terabot's avatar presented and with the Gazepoint GP3 eye tracker under it. In the left photograph of Figure 42, another laptop can be seen in the foreground, which was necessary for real-time analysis of the new system's logs during the dialogues. There, in the right-hand window on the laptop's screen, a real-time preview of the participant's eyes can be seen (thanks to the Gazepoint GP3 software), and in the left-hand window are the dialogue system's logs.

The first step was the design of special scripts for the actors, since the main objective was to simulate the behavior of the patients as closely as possible. It was important to maintain a high quality and keep as close as possible to conversation situations, such as with patients at IPIN. I proposed four different patient behavior scenarios, which were observed during the study at IPIN:

- "positively aroused" – engaged and enthusiastic, sometimes shouting at the system,
- "withdrawn" – shy and anxious, giving only brief answers,
- "neutral" – a little distrustful of the system's operation, but willing to cooperate;
- "challenger" – intentionally inconsistent and sometimes giving bizarre answers.

A detailed description of these scenarios can be found in Appendix B. The participants had to play one chosen role (choose a scenario) and hold one conversation with Terabot on a chosen

a)



b)



**Figure 42:** Experimental setup: “Teatr PW” theater actors conversing with Terabot, a) seen from the side with additional laptop for process control; b) seen from the participant’s perspective

emotion (anger, shame, or fear). After preparations, gaze-enhanced LLM-based dialogue system tests were performed. In total, 20 conversations between the participants and Terabot took place. The conversations lasted between 7 and 10 minutes (depending on utterance length and scenario complication).

## 6.6 Analysis of the new dialogue system performance

Table 7 summarizes the conversations with Terabot. When it comes to the overall performance of the new dialogue system, in all these cases, no situation would lead to a pause in the conversation (as in the experiments with the older version of Terabot, see Chapter 3.3). From the start of the conversation initiated by the participant until the relaxation exercise, all conversations followed the guidelines and had the same order. OpenAI’s safety mechanisms ensured that no inappropriate responses were received during the real-life tests.

Only in the case of conversations according to scenario D, where the participant acted as a “challenger” (see Appendix B, which is based on actual patient behavior), the answers became tricky or did not match at all. In those cases, the conversation did not come to an end, meaning it had to be ended manually in the system. This is because the task was to challenge the system literally, so the participant aimed to give answers that were out of its scope and sometimes contradictory. Sometimes in those cases, no relaxation exercise could be proposed, and no end

to the dialogue was suggested by the system.

These cases can be considered a possibility for future enhancements of Terabot to handle such out-of-scope situations.

### 6.6.1 Eye-tracking response activation

During the conversations, the eye-tracking enhanced response activation took place at the expected moments. It was observed that during the real-time tests, this response activation succeeded 48 times. This is a success for the new dialogue system, since now, using the eye tracker in this real-life scenario, where a dialogue pause was caused, it is indeed helpful and maintains the dialogue flow. During these real-time tests, some situations arose where this response activation was not appropriately activated. These can be divided into two main issues: **too quiet speech** and **interruptions in participants' speech**.

The **first issue** occurred when the participant's speech volume (mostly when playing scenario C, see Appendix B) was very low. In this case, the microphone did not catch the speech signal. After the predefined time, when no speech was detected, depending on the presence of the gaze, an appropriate eye-tracking enhanced activation of Terabot's utterance took place. Since their speech was not detected, some misunderstandings arose for the users. They realized that the dialogue system did not register their answer, so they had to repeat it. It is worth mentioning that the same situation also took place when experiments were performed at IPIN with patients.

The **second issue** occurred when the participants did not answer immediately. When they were thinking about a response and finally decided to answer, in most cases, there was a collision of two signal processing tasks. As already described in previous chapters, when, for example, speech is not detected for 7 s (as described in Chapter 3.3.3), an utterance activation takes place. However, when the participant does not answer for 6 s and then starts to answer, the speech signal processing can be slightly slower than the eye-tracking enhanced utterance activation process. In practice, this means that the participant started answering, but in the meantime, the utterance activation process had started (and was faster than the speech signal detection). In those cases, Terabot's utterance was already activated. This resulted in interruptions of participants' answers by Terabot's utterance. Unsurprisingly, such situations occur, since the topic of proper turn-taking by dialogue agents is a major problem known in the literature [131–133], and also in commercial dialogue agents (e.g. Duolingo).

### 6.6.2 Analysis of conversations

Results of tests on “Teatr PW” theater actors are presented in Table 7. When analyzing the conversations (from the older and the newer version of the spoken dialogue system) the following criteria were applied – a conversation was considered:

- **successful** – if the conversation went smoothly from the beginning to the relaxation exercise;
- **not successful** – when one of the following situations occurred:
  1. the conversation was restarted against the user’s will, or
  2. a Terabot’s response was repeated, or
  3. the flow of the conversation was suddenly and abruptly interrupted (e.g., after the greeting and the agreement to the conversation, a relaxation exercise was abruptly suggested), or
  4. Terabot’s response was completely inadequate and inappropriate for the situation (e.g. failure to recognize irony).

**Table 7:** Results of tests on “Teatr PW” theater actors compared with the results of the older version of the dialogue system

Version of spoken dialogue system “Terabot”	Number of all conversations	Number of conversations with successful flow	No. of conversations with issues other than flow-related
older version (RASA-based)	190	138	89
		5 (only the challenger)	17
multimodal (LLM-based & eye tracker enhanced)	20	16	4
		13 (without challenger’s sc.)	0
		7 (only challenger sc.)	4

It is also important to mention that whenever a problem occurs during a conversation (as indicated by the "No. of conversations with issues other than flow-related" column), this does not imply that the conversation is unsuccessful. Sometimes, even in human-human conversations, minor problems may occur, but the main message is still understood, and the speakers are not uncomfortable or confused. The same logic was used in this case.

Table 7 shows results about successful conversations with the older version of Terabot compared to the new version (LLM-based and eye-tracking enhanced). In the older version of Terabot, the most common problems that occurred during conversations were problems with intent recognition, which sometimes led to disruptions in the dialogue flow or to abrupt restarts. Detailed information about the solutions to these and other problems can be found in the published research articles [134, 135], but will not be explained further, since they are not the aim of the dissertation. When it comes to the analysis of the performance of the new version of Terabot, as the results in Table 7 show, 80% (16/20) of conversations were successful. This percentage was calculated considering all conversations provided by the “Teatr PW” theater actors, regardless of the chosen scenario (whether it was the "challenger" scenario or another). This is a success for the new Terabot version, as it has exceeded the number of successful conversations of the older system (study at IPIN); there, the number of conversations with successful flow was 72.6% (138/190). A deeper analysis shows that when only conversations in the non-challenger scenario were considered (meaning 13 conversations with a non-challenging task were directed to Terabot), all of those succeeded. This is a very promising and encouraging result for the multimodal dialogue system version. The multimodal version of Terabot has been successful, but it has also encountered some issues. The reasons for these incidents were as follows:

**When the user’s utterance was out-of-scope:** this happened when the user chose the “challenger” scenario. In cases where the user’s responses were completely irrelevant to the therapeutic conversation (e.g. asking for advice on washing white sports shoes), and the user posed such questions three times in a row, the LLM produced no suitable answer. Instead, it simply repeated the last utterance given by Terabot, regardless of whether it matched the user’s intent.

**Too loud noises outside of experimental room:** some people were conversing in a loud voice next to the room in which the experiments took place. When the tester talked in a standard or silent voice, the speech volume was probably not as loud as the background noises, which is why sometimes the user’s speech was not recognized appropriately. This resulted in no speech being detected by voice activation detection (VAD) in the ASR module, so the users had to repeat the question.

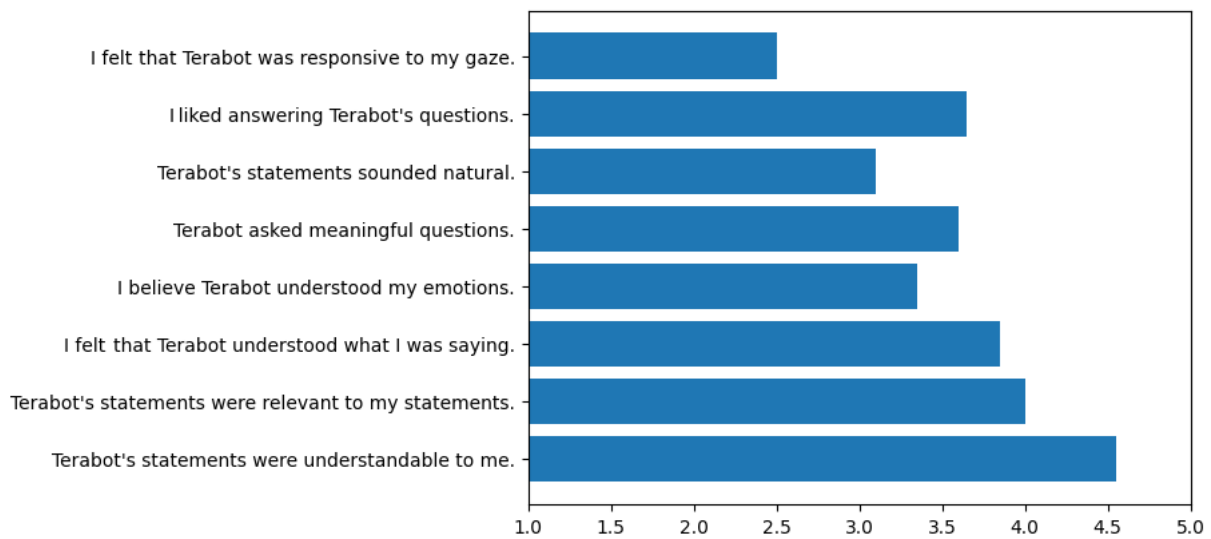
**Speech recognition problems:** this problem only occurred once when a user responded in informal Polish. The person said "no" in response to Terabot’s question, which means "OK" in informal Polish. The Polish "no" was detected and transcribed as the English word "no" (which is a negation). Terabot then interpreted this as the user refusing to answer the question. This caused Terabot to respond appropriately (meaning that the user did not want to participate in the relaxation exercise, etc.), and the dialogue ended. Since this was not the desired behavior, it was perceived as an abrupt ending of the dialogue by Terabot.

**Failure to recognize irony in the user’s utterance:** in one case where the user selected the "challenger" scenario, the person spoke ironically when describing a situation. This irony was not recognized, so the LLM produced an unsuitable utterance and response tone. This happened a couple of times since the user used irony in many subsequent answers.

It is important to mention that these out-of-scope situations were not specified in the rules of Terabot (in the Assistant API or system instructions of Terabot). As this is an issue related to the operation of the LLM (here, the chosen model GPT-4o-mini), it should be studied more deeply. Thus, these situations serve as important indicators when considering future work on Terabot.

### 6.6.3 Survey on participants' satisfaction regarding conversations with Terabot

After conducting the test on the gaze-enhanced LLM-based dialogue system, each participant filled out a survey on different aspects of the system. The results are shown in Figure 43. The data shows that, according to the participants' opinions, Terabot's statements were understandable (mean value = 4.5/5). They also state that they felt understood while conversing with the dialogue system. Terabot's statements were received as relevant (appropriate) to the participant's utterances (mean value approximately 4/5). These promising outcomes show that users responded positively to the new LLM-based, eye-tracker-enhanced system.



**Figure 43:** The “Teatr PW” theater actors’ satisfaction regarding dialogues with LLM-based Terabot

Participants reported that Terabot sounded natural (mean value = 3/5); but in additional descriptive opinions, many of them stated that they dislike the dialogue system’s voice. The reason might be due to too high expectations towards Terabot, since, in general, the dialogue system provided the actors with successful, meaningful conversations. Changing Terabot’s voice might be one of the factors worth considering when thinking about further research.

When comparing the satisfaction results of the actors with those of patients at IPIN (see Chapter 3.4, Figure 27), it can be concluded that both groups found Terabot’s utterances understandable. It is also important to note that both groups found Terabot asked meaningful

questions. A comparison of the satisfaction results of these two groups shows that both patients and actors highly rated Terabot in terms of emotional understanding (average score was around 3.2). To sum up, these satisfaction outcomes are a very encouraging and promising result in terms of the further development of Terabot.

## 6.7 Conclusion

The offline tests with data from conversations at IPIN show that the new version of the spoken dialogue system, which is gaze-enhanced, would help in situations where the dialogue is set to pause. This has been described in the research article [136]. The real-life tests with the “Teatr PW” theater actors showed that the new dialogue system was well received, with its statements being understandable and relevant. This is a very important and promising result because it demonstrates the potential for future research regarding Terabot.

There are several aspects worth changing (e.g., the voice of Terabot), especially considering the testing of this dialogue system on patients with psychiatric disorders. What is also worth thinking about is the collision of the speech signal processes and the gaze-based utterance activation process; there might be potential for optimizing these processes. Another important topic not to forget is the safety of Terabot’s responses, considering that they are LLM-based. An important next step should be an appropriate “response safety check filter” applied to all responses received by the LLM, before they are spoken out by the TTS module to the user.

To summarize, the use of the eye tracker strongly improved the dialogue performance of Terabot. The tests on the gaze-enhanced and LLM-based dialogue system prove that by using the eye tracker, most of the problems identified during conversations at IPIN could be solved. This means that **the research objective “[RO3]: Enhancement of fluency of conversations with a dialogue system by using an eye tracker” has been achieved.**

# Chapter 7

## Automatic assessment of patient engagement

This chapter will address the issue presented in Chapter 3.3.4, which is the improvement of the dialogue system with a functionality that would provide information about patients' presence and engagement in the conversation.

### 7.1 Problem description

During the study at IPIN, an assistant was needed each time a patient wanted to talk to Terabot. The assistant's task was to note whether the patient was engaged in the conversation with Terabot (the patient's willingness to talk, concentration during the dialogue, etc.) or if there were any unforeseen situations (e.g. if the patient suddenly stopped interacting with Terabot). The psychiatrist was interested in how the conversations went (information gained through transcriptions of the patient-Terabot conversation) and the overall behavior and engagement in the dialogue. These two components were important for properly evaluating the patient's progress during this study. However, this approach was time-consuming. The idea is to use tools that could be helpful and provide relief for people. This approach was also prone to human error. Although the main purpose of their notes was clear to all assistants, they were written in different ways, from various perspectives, and did not always cover the same topics. Various assistants reported the conversations in a variety of ways. This was uninformative for properly evaluating the therapeutic sessions.

## 7.2 Analysis of assistants' notes on patient engagement during dialogues

**Table 8:** A sample of assistants' comments provided after patients' therapy sessions with Terabot, a direct translation of the notes from the Polish language.

Assistant ID	Patient ID	Provided note
A1	P01	At first, he spoke quietly, Terabot 3 times did not respond; "what would your body want to do if anxiety took control?" - "I do not know"; after the exercise - "how do you feel now?" - "I do not know"; "did the exercise help?" - "no"; when asked after the session, he states that he finds it hard to recognize what is happening to him when he feels anxious and does not need the exercise now.
A2	P01	visible symptoms of anxiety, tension
A3	P02	(session 1) The patient is a bit skeptical, worried that the conversation will be like with phone bots. After talking to Terabot, he says "it was not so bad, he even understood me." He speaks clearly, at normal volume. There are no problems during dialogue - everything is understandable.
	P02	(session 2) The patient is a little frustrated that Terabot offers the same exercise and expects something different despite the same choice of exercise emotions. He expected a deeper dialogue about his problem. Responds vividly, in full sentences, sometimes impatient. The patient is eager to talk and has a lot of energy. Responds in full sentences and wants to go deeper into the topic.
	P02	(session 3) Patient is cooperative, willing to answer Terabot's questions, and delves into what ails him during dialogues. Goes into detail during dialogues. He has many suggestions for improvement and wants to get involved himself. He talks about his problems even after talking to Terabot. He is open, full of energy, and ideas.

<b>Assistant ID</b>	<b>Patient ID</b>	<b>Provided note</b>
A4	P03	(session 1) The patient is rather cautious toward Terabot, tending to respond only with incomplete sentences or yes/no.
	P03	(session 2) Patient patient, even when Terabot asks for a repetition he tries to describe it differently. He tries to perform the exercise reliably.
	P03	(session 3) The patient is already answering the Terabot's questions a bit more openly. Moderate voice volume.
A5	P04	(session 1) Eager to work with Terabot, calm, works with enthusiasm, Rating how much my presence interfered (0-100): 0.
A6	P04	(session 2) After being asked, if she remembered feeling shame she answered "yes", the Terabot did not respond for a long while, she had to repeat "yes" - then he responded, during the session the patient waited a long time for the Terabot to respond also after her other answers, she did not do the homework because she "did not feel anger yesterday".
A5	P04	(session 3) The woman brought the worksheets. Calm, composed, attentive and focused on the session with Terabot throughout. Eagerly completed the final questionnaire, Rating how much my presence interfered during the session (0-100): 0.

Table 8 presents some of the assistants' notes. The names of patients and assistants have been changed into individual ID numbers. For example, the provided notes describe a situation in which Patient P01 was seen by Assistant A1 one day and Assistant A2 on another day. Analyzing the notes of both assistants, it can be seen that the first note is detailed, contains quotations of the patient's sentences, describes the patient's behavior, and the volume of the patient's voice. The second note is very concise and describes the patient's behavior/mood only very briefly.

In summary, all the notes presented in Table 8 show that the different notes taken by the assistants cover different aspects. Despite a commonly stated goal, different assistants pay attention to different aspects during the patient-Terabot conversation. It can also be seen that the focus of the notes is quite different and that the style of notes varies from person to person. It is clear that each assistant considered different information to be important. When trying to

find some commonly mentioned topics, they can be divided into several categories:

- information on the volume of the patient's speech;
- information on the complexity of the sentences (sometimes a direct transcription of the patient's sentences);
- assessment of the patient's involvement during the dialogue or during a relaxation exercise;
- assessment in the form of marks on whether the presence of an assistant is disturbing (patient's feelings);
- information on whether the patient completed the questionnaire on the last day;
- information on whether the patient completed a homework assignment (i.e. filling in a sheet describing the situations and overwhelming emotions experienced during them);
- others, not mentioned in the example notes.

The provided samples of notes do not present objective, standardized data/information from patient–Terabot conversations. Since information about the patient's behavior is necessary for the evaluation of the therapy sessions, a solution is needed that would provide clear, objective results based on the data that can be easily compared (for all patients).

### **7.3 Proposed method**

The main goal was to find an objective way to measure patient engagement. For this purpose, I came up with the idea of using eye-tracking data. Authors of the research paper [137] suggest that the exploration of different parameters (e.g., ECG, EEG) and the combination of them allows user interfaces to collect information and distinguish between four attentional states of the user while interacting (at rest, moving, thinking, and busy). Based on this idea, I came up with the concept of combining audio data with eye-tracking data to gain insight into the patients' engagement during dialogues. Based on the experiences at IPIN, an engagement matrix was proposed, which would enable differentiation between the most distinctive patients' behavior. The concept of collecting and processing data from patient-Terabot conversations is schematically illustrated in Figure 44.

After the patient has spoken to the dialogue system, a general assessment in the form of a note could be automatically generated based on the speech/transcription and eye-tracking data collected during the conversation. The concept of the patient's engagement matrix consists of analyzing two signals: the transcriptions of speech data and eye-tracking data. From the transcriptions, the text can be analyzed, giving an idea of how open the patient was while



**Figure 44:** Data collected to determine patient involvement during dialogues with Terabot

conversing. Concise answers like “yes/no” would indicate low engagement; while extensive utterances with many intents indicate unnaturalness and/or even the patient’s verbosity. Additionally, eye-tracking data would be used to identify gaze contact with the dialogue agent, which is considered an important factor in therapeutic sessions (as already pointed out in Chapter 5.5). For example, if during the conversation there was very little or no eye contact (meaning the patient’s gaze was not aimed at Terabot), this information should be noted to inform the psychiatrist.

As a result of the engagement matrix might help to distinguish between the following states: if the patient is involved in the dialogue versus the patient not being involved (fewer words provided); and if eye contact is maintained versus no eye contact present during the conversation. It could therefore attempt to automatically gather the patient’s involvement in the dialogue and analyze the progress in therapeutic conversations.

## 7.4 Results

To get an objective assessment of the proposed solution, the transcriptions and eye-tracking data of IPIN’s patients were used. Some patients’ data could not be analyzed due to data collection problems (which have already been described in previous chapters). All the data that it was possible to collect was analyzed. Looking at the results collected in Table 9, it can be seen that different patients show different types of objective fixation data.

The data of Patient 04 indicates that this person experienced verbosity and had a significantly longer average speech length than the other patients. Additionally, the eye-tracking data could not be recorded due to the patient’s behavior. When comparing the fixation data between the patients, one can see that the fixation duration times may differ (at Terabot’s person, at the screen, or outside of the screen). Therefore, I suggest that the objective measures (e.g., fixation data on Terabot’s person) should not be interpreted in terms of comparing numbers. In this case, the zero-one information is the most appropriate, i.e. whether or not the patient’s gaze was recorded in these three individual categories (Terabot’s person – screen – outside of screen). If so, it means that eye contact was maintained during the therapeutic dialogue. If there was no record of their gaze in any category or only their gaze outside of the screen or outside

Engagement in the conversation		
	Low	High
Type of signal	speech - short utterances (a few words, e.g., max. 3 words) - volume (low to moderate), - speech rate (low/very high)	- longer utterances (>5 words) - volume (moderate to high), - speech rate (normal) - special case? ***
	gaze - almost no eye contact, (mostly looking away from Terabot's interface)	- normal/intense eye contact with Terabot's interface - special case? ***

\*\*\* Indicator of unnaturalness: very vast, long utterances + very intense eye contact (starring at screen)

**Figure 45:** The proposed patient's engagement matrix based on speech and gaze

**Table 9:** Information collected on patients' eye-tracking and speech data.

Patient ID: Short (subjective) patient description	Fixation duration – Terabot's person	Fixation duration – screen (not Terabot's person)	Fixation duration – outside of the screen	Avg. word count per utterance
01 co-operative behavior	3.25 s	0.52 s	10.5 s	3.8
02 not talkative	6.84 s	0.6 s	14.5 s	1.2
03 focused	10.29 s	0.2 s	5.3 s	1.0
04 verbosity	-	-	-	9.7

of Terabot’s person, then a psychiatrist would need to examine this patient and this behavior more deeply. This information, together with the average word count per utterance, can give information about whether the patient was engaged. Table 10 presents the solution of applying the engagement matrix to the collected audio and eye-tracking data. Figure 46 shows a proposed result in the form of notes, which could be given to the psychiatrist or psychotherapist.

**Table 10:** Result based on the proposed patient engagement matrix on patients’ data.

<b>Patient ID: Short (subjective) patient de- scription</b>	<b>Fixation duration data – Terabot’s person</b>	<b>Fixation duration data – screen (not Terabot’s person)</b>	<b>Fixation duration data – outside of the screen</b>	<b>Avg. word count per utterance</b>	<b>Result</b>
01 co-operative behavior	✓	✓	✓	3.8	eye contact involved
02 not talkative	✓	✗	✓	1.2	eye contact few words
03 focused	✓	✗	✓	1.0	eye contact few words
04 verbose	✗	✗	✗	9.7	no eye contact, many words

For a more in-depth and detailed analysis of the fixations during dialogues, a number of other factors should be taken into account, e.g. the summed duration of blinks, time without detection of the eyes, etc. Then it would even be possible to give the percentage of time looking at the Terabot’s AOI or the total time beyond fixation on the screen. Furthermore, one must bear in mind the difficulties in eye-tracking data collection at IPIN and that such a detailed analysis is very challenging. Again, when considering the purpose of analysis, bearing in mind that the aim is to provide only general information about the patient’s behavior while conversing, such an in-depth analysis might seem overly detailed and unnecessary. This is why the result proposed in Table 10 is considered to be fully sufficient.

Patient 01: The patient was **involved in the dialogue** (spoke with an average word count of approximately 4 per utterance). The patient **maintained eye contact** during the conversation with Terabot. The person looked at Terabot's person and also outside of the screen while conversing.

Patient 02: The patient was **not really involved in the dialogue** (spoke with an average word count of approximately 2 per utterance). The patient **maintained eye contact** during the conversation with Terabot. The person looked at Terabot's person and also outside of the screen while conversing.

Patient 03: The patient was **not really involved in the dialogue** (spoke with an average word count of approximately 2 per utterance). The patient **maintained eye contact** during the conversation with Terabot. The person looked at Terabot's person and also outside of the screen while conversing.

Patient 04: The patient was **very involved in the dialogue** (spoke with an average word count of approximately 10 per utterance).

INDICATION: This is **worth analyzing deeper**.

The patient **did not maintain eye contact** during the conversation with Terabot. The person did not look at Terabot's person, nor outside of the screen, while conversing.

INDICATION: This is **worth analyzing deeper**.

**Figure 46:** Proposed result in form of note about each patient's engagement for the psychiatrist

## 7.5 Conclusion

The proposed solution shows that using an eye tracker in a dialogue agent in clinical practice would allow the collection of identical information characteristics for each conversation. In terms of limitations, it has to be stated that the proposed solution was based on the available eye-tracking data of only a few patients. To confirm the validity of the proposed approach for patient engagement assessment, this solution should be tested on a larger eye-tracking data and audio dataset.

Nevertheless, the proposed approach is resistant to human errors. It would provide easily accessible information for every psychotherapist and psychiatrist interested in patient engagement during therapeutic sessions. There would be no need to do a full analysis of the conversation unless psychiatrists request it. It would save not only the valuable time of the medical specialists but also costs (there is no need to hire additional staff to be present during the therapy and write notes).

As a result of this solution, the therapy sessions with Terabot do not require additional support. Any patient who is in need of a conversation can sit in front of the dialogue system. These conversations, whenever they take place, can still be recorded and provide the psychiatrist with information about the engagement of the patient. This approach makes the dialogue system more accessible to patients whenever they want to talk, and not whenever an assistant is available. The results of this chapter have been presented at the Third Digital Mental Health Conference in Grenada, Spain. It can be stated that **the research objective “[RO4]: Enabling automatic user engagement assessment while interacting with the dialogue system” has been achieved.**

# Chapter 8

## Summary and discussion

In this work, I focused on analyzing how the use of eye tracking can improve a therapeutic dialogue system. I addressed several problems that arose during the study performed at IPIN.

The first main issue was whether the eye tracker could be used during the dialogues with Terabot without prior calibration (Chapter 4). Since the study's main aim was to provide the patients with therapeutic conversations, it was impossible to perform a calibration beforehand, as this could have discouraged the patients from opening up and disrupted their comfort. After conducting experiments, the results showed that the average pupil size values and their corresponding confidence values indicated no difference in the data received with and without calibration. Validating this hypothesis would require a larger study and confirmation through statistical tests. Large-scale tests with uncalibrated data have been performed by other research groups, e.g. [105, 138, 139]. Their results indicate that calibration is not always needed. Results of the tests in my study, although conducted on a small number of participants, strongly suggest a high probability that calibration of the Gazepoint GP3 eye tracker is not needed. This means that the eye tracker could be used with a dialogue system without prior calibration. In this way, the research objective **“[RO1]: Investigating whether an eye tracker needs to be calibrated when used in a spoken dialogue system”** has been successfully achieved.

During the experiments at IPIN, some issues arose when patients conversed with Terabot. These issues caused the dialogue to pause, making patients feel uncomfortable. The first case occurred when Terabot interrupted the patients' answers. Patients gave answers that were sometimes very short or sometimes very long, with many pauses during their responses. The dialogue system did not recognize the pause appropriately, detecting it as the end of the utterance and proceeding with the dialogue flow. Such problems, known from research articles on spoken dialogue systems (e.g. [28]), and termed as 'endpointing' or 'endpoint detection', are very challenging. In this case, when the patient had not finished speaking, Terabot's utterances interrupted the rest of the patient's answer.

The next issue occurred when Terabot made the patients wait for its reply for a very long time. When patients' utterances were too quiet or too rapid, the ASR module could not detect them as valid speech input. The microphone was still on, and the patients did not realize that

their responses were not being recognized. This situation unintentionally paused the flow of the conversation. Such waiting made patients feel insecure or even irritated. On the other hand, there was also another issue: there were long periods of time where the patients did not respond to Terabot. During this time, the microphone was still on, waiting for a speech signal to be detected. The patients explained after the conversation that they wanted to think about the answer deeply. Sometimes they reported that their thoughts went somewhere else, and it took them time to come back to the conversation. In these cases, it was found that non-verbal information (collected by the eye tracker) could be beneficial for the dialogue system, because it could give non-verbal information about patient behavior in those cases. In order to solve this problem, patients' fixation data from the dialogue system's visual interface while they were speaking the closing phrases of their utterances were analyzed. The results showed that while finishing speaking, patients' fixations were directed at Terabot's person. This provided a solution that could be incorporated into the dialogue system. Terabot's utterances could be activated when the patient's utterance ends, thus speech had been successfully detected (and come to an end), and at the same time, fixations were detected at Terabot's person. Therefore, **the research objective “[RO2:] Analysis of areas of interest in the graphical interface of the dialogue system” has been achieved.**

Although this kind of mechanism of gaze-activated responses seems helpful for indicating the closing phrases of patients' utterances, there may be a significant risk. While supported by the data analysis results, this solution could potentially make Terabot less accessible to patients with schizophrenia. Since the aim is to make Terabot accessible to all patients in need, I decided not to implement this solution to the dialogue system. In therapy, eye contact is important, but it should not be considered important for the functionality of a dialogue system for therapeutic purposes.

The next step consisted of designing and implementing a gaze-enhanced dialogue system, which considered the previous results. Since research articles [57] provide evidence that LLMs are increasingly being used for dialogue management, the new version of the dialogue system was LLM-based. I used one of OpenAI's GPT models, with Assistant API integration. The new multimodal dialogue system showed that real-time eye-tracking data can indeed serve as an additional input into a dialogue system, which enhances its functionality and improves dialogue fluency. As a result, **the research objective “[RO3]: Enhancement of fluency of conversations with a dialogue system by using an eye tracker” has been achieved.**

To verify the performance of the new multimodal dialogue system, offline tests with patients' data and real-time tests on users were carried out. The offline tests showed that the multimodal dialogue system could activate more situation-appropriate responses when patients' data from IPIN was used. The real-life tests with “Teatr PW” actors demonstrated that real-time eye-tracking data could support the dialogue in situations where no speech was detected in authentic situations. These tests also showed that the new dialogue system was well-received, with understandable and relevant statements. This was a very important result, as it demonstrates that

eye-tracking can enhance the dialogue flow.

The last issue that I addressed was that during the conversations there was no non-speech related information about patient participation during relaxation exercise or conversation. In the RASA-based version of the dialogue system, there was no means of contacting the dialogue agent other than speech. This means that there is no place for any feedback from the patient during the relaxation exercise. Since no speech is collected, there is no certainty that the patient is even sitting in front of the screen and listening, or maybe has just walked away. I proposed the idea of an automatic assessment of user engagement while interacting with the dialogue system using eye-tracking data and audio data in the form of the patient engagement matrix (in Chapter 7). This matrix concept involves analyzing speech transcriptions and eye tracking. The text from the transcriptions can be analyzed to give an idea of how open the patient was during the conversation. Additionally, eye-tracking data is used to identify gaze contact with the dialogue agent, an important factor in therapeutic sessions. As a result of the matrix, a short note is generated for the psychiatrist, which contains the most important information (if there was eye contact with the patient during the dialogue and how talkative the person was). This solution provides easily accessible information for any psychotherapist or psychiatrist. There is no need to conduct a full analysis of the conversation, which saves the time of medical professionals and reduces costs. It can be stated that **the research objective “[RO4]: Enabling automatic user engagement assessment while interacting with the dialogue system.” has been achieved.** It has to be mentioned that the solution is based on the available eye-tracking data of only a few patients. To confirm the approach’s validity, it should be tested on a larger eye-tracking and audio dataset.

Finally, it is important to mention the key accomplishments through which the above-described research objectives were achieved:

1. Performing preliminary eye-tracking tests on the topic of the necessity of calibration;
2. Performing extensive testing of the Terabot dialogue system;
3. Providing assistance at IPIN during dialogues between patients and Terabot;
4. Performing post-processing and analyses of collected eye-tracking data and ASR results;
5. Collecting a database of eye-tracking data combined with audio data (patient-Terabot conversations);
6. Implementing a multimodal version of the dialogue system (speech-to-speech): with an LLM-core (GPT model by OpenAI) and enhanced with real-time eye-tracking data;
7. Performing real-time tests of the new version of Terabot on the “Teatr PW” theater group.

In summary, the issues encountered with the spoken dialogue system have been successfully solved using eye tracking. This means that **the aim of this thesis “The use of eye-tracking can improve a therapeutic spoken dialogue system” has been fulfilled.**

## 8.1 Limitations

Some issues need to be discussed in view of potential limitations. Whenever performing eye-tracking experiments, especially with stationary eye trackers, it is advised that the participant's head should be kept as still as possible. This is important to ensure that high-quality data is collected. However, as this could greatly affect the comfort of patients, it was not a requirement for the patients at IPIN. The freedom of movement during the therapeutic dialogue was of the utmost importance. This is the reason why some patients were moving (head or body) while talking to Terabot. As already described in previous chapters, the reasons for that were that, for example, the patients were at different stages of their illness, and some could not sit without moving, while others sat in an uncomfortable position and remained in it through the whole conversation (with fixed fixation outside of the screen). They also had different individual concentration levels during the conversations (which was shown in the body movements while conversing). This is why it was difficult and sometimes impossible to obtain more good-quality data from the eye tracker. The analysis of the experiments (e.g. those presented in Chapters 5 and 7) was based on a smaller amount of data than the entire collected eye-tracking and audio dataset.

Another idea for the use of eye-tracking data for the spoken dialogue agent was to measure if they really felt relaxed during the relaxation exercise. This could have been analyzed by having the pupil diameter data from relaxation exercises, with the data prepared accordingly before it (e.g., measuring the baseline of each patient). But since not every patient could or even wanted to perform the relaxation exercise, and because of their movements, collecting good-quality pupil data from patients with schizophrenia became impossible. Again, this is the reason why only some particular eye-tracking data was used for analysis.

This dissertation illustrates the challenges engineers must consider when developing HCI solutions for individuals with mental health disorders. As can be seen, there is still much work to be done in the area of computer-assisted digital mental health.

## 8.2 Further work

Despite all the work that has been done, it is important to note that there is still room for further improvement of Terabot. The following possibilities can be considered.

In terms of average fixation times of Terabot's interface, a closer look should be taken at the "AOI buttons" area (mainly the "Reset" button and the microphone symbol). The average and maximum fixation durations were comparatively high. The reason for this might be the changing color of the microphone symbol during the conversation (it was black when Terabot was talking with the microphone off, and it turned red when the microphone was on to record the patient's speech). During the conversations, as some patients discovered that the microphone symbol changed colors, their fixations did not come back to Terabot's face at all; it attracted their

gaze till the end of the conversation. Because some patients were gazing at these areas, it should be considered adjusting the colours so as not to attract the patients' unnecessary attention. This information is important when thinking about further development of Terabot's interface.

There are also other topics worth exploring further, particularly when the aim is to test this system again on patients with psychiatric disorders. As already mentioned in Chapter 6.7, the multimodal activation between the speech signal processing and gaze-based utterance activation should be analyzed more deeply. There may be potential to optimize these processes, so no collision will take place.

Another important topic is the safety of Terabot's responses, considering that they are LLM-based. Recently, there have been press reports about dangerous incidents involving the use of chatbots based on LLMs, like ChatGPT [140]. These incidents demonstrate that AI is advancing and becoming more widely used by private individuals. This is why scientists and engineers focusing on LLMs must prioritize making LLMs safe and usable by the general public. Fortunately, LLM companies (such as OpenAI) began to look into it more thoroughly [141]. They report that they are improving how the GPT models respond during sensitive interactions. They state that they are working on specific safety enhancements, particularly regarding emotional dependency and mental health crises. Nevertheless, in the case of Terabot, an appropriate "response safety check filter" should definitely be applied to all responses received by the LLM before they are spoken by the TTS module. This is now done automatically by OpenAI's safety filters, which is why no inappropriate answer was received during the real-life tests. However, if this dialogue system is to be used in a psychiatric hospital again, it would be important to implement its own filter and carry out further checks to ensure that patients cannot experience any harm when interacting with Terabot.

During the tests, it was also recognized that whenever two or three subsequent questions (posed by the participant) were out of scope, then the system would just reproduce the last given answer, regardless of the intent of the questions asked. As this is an issue related to the operation of the LLM itself, it definitely should be studied in more depth when considering further research with Terabot.

After applying appropriate adjustments to Terabot, the dialogue system may be ready for the next step: providing help during pilot clinical trials with patients with psychiatric disorders.

# Appendix A

## Dialogue sample – chosen topic: anger

**Table 11:** Sample dialogue on anger between a patient and Terabot (in the Polish language).

Pacjent	Terabot
Cześć!	
	Dzień dobry! Jakie ćwiczenie dziś wykonujemy?
Poproszę ćwiczenie na temat złości.	
	Świetnie, witam Cię, pacjencie! Nazywam się Terabot i pomagam w Twojej terapii. Chciałbym z Tobą dziś przeprowadzić ćwiczenie dotyczące złości. Czy wyrażasz zgodę, abyśmy porozmawiali?
Tak, spoko.	
	Dziękuję za Twoją zgodę. Przypomnij sobie jakieś zdarzenie, które wywołało u Ciebie złość. Być może coś takiego się zdarzyło w domu, w pracy albo na ulicy. Czy przypominasz sobie coś takiego?
No pamiętam...	
	Pamiętasz, gdzie to się zdarzyło?
Podczas zakupów w sklepie.	
	To prawda, przy zakupach łatwo o nerwowe sytuacje. Czy denerwujesz się nagle, czy wybuchasz?
Zazwyczaj wybucham w takich sytuacjach.	
	Zauważ, że złość bardzo szybko narasta. Wtedy twoja uwaga się zawęża. Koncentrujesz się na tym, co Cię złości. Z kolei ignorujesz wszystko, co normalnie mogłoby Cię zainteresować. Na przykład, jeśli złościmy się na partnera, to zapominamy o tym, jak bardzo go kochamy. Przypomnij sobie, jak odczuwasz złość w ciele. Jak Twoje ciało wtedy reaguje?
Czuję, że serce mi wali jak młot.	
	....

**Table 12:** Sample dialogue on anger between a patient and Terabot (in the English language).

Patient	Terabot
Hi!	
	Good morning! What exercise are we doing today?
I'd like an exercise on anger, please.	
	Great! Welcome! I'm Terabot, and I'm here to help with your therapy. Today, I would like to do an exercise with you about anger. Is it all right if we talk about it?
Yes, it's ok.	
	Thank you for your consent. Think back to an event that made you angry. Perhaps something like this happened at home, at work, or on the street. Can you remember something like that?
I do remember...	
	Do you remember where it happened?
While shopping at a store.	
	It's true, shopping can easily lead to stressful situations. Do you suddenly get nervous or do you explode?
I usually explode in situations like this.	
	Notice how quickly anger builds up. Then, your attention narrows. You focus on what makes you angry. Consequently, you ignore everything that would normally interest you. For example, when you're angry with your partner, you forget how much you love them. Remember how anger feels in your body. How does your body react?
My heart pounds like a hammer.	
	....

# Appendix B

## Scenarios for actresses and actors of “Teatr PW” theater

### **In English:**

Acting assignment description:

#### PATIENT IN SESSION WITH VIRTUAL THERAPEUTIST

Your mental state has not been the best lately. You are admitted to a psychiatric hospital, where, in addition to drug treatment, you take part in therapeutic classes with psychiatrists and psychotherapists. As part of your treatment, you also have the opportunity to talk to a virtual therapist called Terabot, who can help you work through overwhelming emotions such as anger, anxiety and shame.

- **Behavior type A (“neutral”)**

Initially, you are somewhat distrustful of Terabot and the system’s operation, but despite this, you are willing to cooperate. You respond in a normal voice, sometimes using short sentences and sometimes long ones. You pause for a moment to consider more difficult questions, looking away from the monitor. You do not cause any problems during the dialogue and take part in a relaxation exercise.

- **Behavior type B (“positively aroused”)**

Although you are apprehensive at first, you quickly become very engaged and enthusiastic about the system. You shout when Terabot does not understand something. You generally respond loudly, and your statements are short. You are a patient with a lot of inner tension, which is evident in your expressiveness.

- **Behavior type C (“withdrawn”)**

You approach Terabot shyly and anxiously, and this is apparent in every word you say. You only respond briefly. Sometimes you stare at the screen for a long time without answering, thinking for a long time. Participating alone is exhausting; you are afraid of how the conversation will go. You become too scared and give up on the exercise. You experience moments of withdrawal and internal tension.

- **behavior type D (“challenger”).**

You test the system, sometimes providing inconsistent and sometimes bizarre answers. For example, you choose an exercise on anger, but when answering questions related to anger, you say that you do not have any problems with it. You are surprised by Terabot’s questions. Sometimes, you do not answer at all for a long time, looking away while doing so, e.g., not answering at all.

## **In Polish (original):**

Opis zadania aktorskiego:

### **PACJENT PODCZAS SESJI Z WIRTUALNYM TERAPEUTĄ**

Twój stan psychiczny nie jest ostatnio najlepszy. Trafiasz do szpitala psychiatrycznego, aby tam podczas dłuższego pobytu, oprócz leczenia farmakologicznego, brać udział m.in. w zajęciach terapeutycznych z psychiatrami i psychoterapeutami. Dodatkowo podczas Twojej terapii rozmawiasz z wirtualnym terapeutą o nazwie “Terabot”, który może pomóc przy przepracowaniu przytłaczających emocji takich jak np. złość, lęk, wstyd.

- **Typ zachowania A (“neutralny”)**

Początkowo podchodzisz z pewną nieufnością do rozmowy z Terabotem oraz działania tego systemu, ale mimo wstępnego oporu chętnie z nim współpracujesz. Odpowiadasz z normalną głośnością, używasz czasem krótszych, czasem dłuższych zdań. Zastanawiasz się dłuższą chwilę nad trudniejszymi pytaniami, odwracając wzrok od monitora. Nie stwarzasz żadnych problemów podczas dialogu, bierzesz udział w ćwiczeniu relaksacyjnym.

- **Typ zachowania B (“pozytywnie pobudzony”)**

Masz duże obawy na początku, ale jesteś bardzo pobudzony/-a oraz entuzjastycznie nastawiony/-a do systemu. Krzyczysz, gdy Terabot czegoś nie rozumie. Ogólnie odpowiadasz bardzo głośno, Twoje wypowiedzi są raczej krótkie. Jesteś pacjentem o dużym wewnętrznym napięciu, ujawnia się to poprzez dużą ekspresyjność.

- **Typ zachowania C (“wycofany”)**

Podchodzisz do Terabota z dużą nieśmiałością, lęk jest wyczuwalny w każdym wypowiedzianym słowie. Odpowiadasz tylko zdawkowo. Czasami długo się zastanawiasz, wpatrując się w ekran przez dłuższy czas, nie udzielając odpowiedzi. Sam udział w badaniu kosztuje Cię dużo energii, obawiasz się, jak rozmowa będzie wyglądać. Zbyt się boisz i rezygnujesz z ćwiczenia. Doświadczasz momentów wycofania i wewnętrznego napięcia.

- **Typ zachowania D (“challenger”)**

Testujesz system, dając czasem bardzo dziwne, niespójne odpowiedzi. Np. wybierasz ćwiczenie na temat złości, ale podczas odpowiadania na pytania związane ze złością, odpowiadasz, że nie masz z nią problemów. Dziwisz się, słysząc pytania Terabota. Czasem nie odpowiadasz wcale przez długi czas i odwracasz przy tym wzrok.

# List of Figures

1.	Types of delivery of interventions across different types of digital tools.....	3
2.	Burden of mental health diseases (2021) worldwide and in Poland .....	12
3.	Basic structure of a dialogue system .....	14
4.	Snapshots of the Woebot conversational agent (chatbot) on a mobile phone .....	17
5.	A snapshot of the Tess conversational agent (chatbot) on a mobile phone. ....	18
6.	A snapshot of the Wysa conversational agent (chatbot) on a mobile phone.....	19
7.	A snapshot of the Therabot conversational agent (chatbot) on a mobile phone.....	21
8.	Schematic representation of the eye tracker workflow .....	23
9.	Stationary eye tracker model: Tobii Pro X2-30 by Tobii.....	24
10.	Stationary eye tracker, model: EyeLink 1000 Plus by EyeLink .....	24
11.	Eye-tracking glasses (mobile eye tracker), model: Pupil Neon by Pupil Labs.....	25
12.	Integrated eye tracker into an AR/VR system .....	25
13.	Sampling frequencies and the corresponding data parameter measurement .....	26
14.	Illustration of accuracy and precision of an eye tracker .....	27
15.	Illustration of saccades, fixations and microsaccades .....	28
16.	An image of Terabot's visual interface.....	34
17.	RASA-based Terabot system .....	35
18.	A tester sitting in front of Terabot's interface.....	37

19.	Sample of a dialogue between the patient and Terabot.....	38
20.	Photo of the stationary Gazepoint GP3 eye tracker used in the study (with the biometrics toolkit also presented) .....	39
21.	Experimental setup at the IPIN .....	40
22.	Dialogue system (Terabot) response time to patient utterances - ideal situation .....	41
23.	Errors in the dialogue flow for the dialogue system – interrupting patient’s speech .	42
24.	Errors in the dialogue flow for the dialogue system – too long waiting time .....	43
25.	Prolonged waiting for the patient’s answer .....	43
26.	Lack of feedback for dialogue system .....	44
27.	Patients’ satisfaction regarding dialogues with Terabot .....	45
28.	Experimental setup at the WUT .....	48
29.	Experiment scheme with and without calibration .....	48
30.	Violin plots for right pupil diameter data .....	51
31.	Average values of participants’ right pupil diameters with their corresponding confidence values .....	52
32.	Example of a patient’s utterance and the closing phrase .....	55
33.	Photo of Terabot’s interface with selected Areas of Interest (AOIs).....	55
34.	Visualization Terabot with color intensity .....	57
35.	Scheme of possible response activation indicator .....	58
36.	Probability density bars of waiting times .....	61
37.	Schematic representation of the new, multimodal dialogue system .....	63
38.	Gaze behavior – case A .....	64
39.	Gaze behavior – case B .....	65

40.	Gaze behavior – case C .....	65
41.	Gaze-response decision logic, where X – waiting threshold [s] .....	66
42.	Experimental setup: “Teatr PW” theater actors conversing with Terabot .....	69
43.	The “Teatr PW” actors’ satisfaction regarding dialogues with Terabot .....	73
44.	Data collected for patient engagement analysis .....	79
45.	Proposed patient’s engagement matrix .....	80
46.	Proposed result in form of note .....	82

# List of Tables

1.	Various digital technologies based on [3] .....	2
2.	Eye-tracking parameter groups. ....	30
3.	Inclusion and exclusion criteria. ....	37
4.	Eye Tracker Technical Specifications .....	41
5.	Fixation durations in AOIs.....	56
6.	Waiting threshold values .....	67
7.	Results of tests on “Teatr PW” theater actors .....	71
8.	Sample of assistants’ comments. ....	76
9.	Information collected on patients’ eye-tracking and speech data. ....	80
10.	Result based on the proposed patient engagement matrix on patients’ data. ....	81
11.	Sample dialogue on anger (in Polish). ....	89
12.	Sample dialogue on anger (in English). ....	90

# Bibliography

- [1] Z. Dongjun, W. Mingyue, L. Xinqi, W. Lina, W. Jiali, and J. Mengyao, “Trends in depressive and anxiety disorders among adolescents and young adults (aged 10–24) from 1990 to 2021: A global burden of disease study analysis,” *Journal of Affective Disorders*, vol. 387, p. 119491, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0165032725009334>
- [2] The Guardian. (2024, Nov.) One in three consultant child psychiatrist posts in england are vacant, analysis shows. Accessed: 2025-07-03. [Online]. Available: <https://www.theguardian.com/society/2024/nov/28/one-in-three-consultant-child-psychiatrist-posts-in-england-are-vacant-analysis-shows>
- [3] R. R. Bond, M. D. Mulvenna, C. Potts, S. O’Neill, E. Ennis, and J. Torous, “Digital transformation of mental health services,” *npj Mental Health Research*, vol. 2, no. 1, p. 13, 2023. [Online]. Available: <https://doi.org/10.1038/s44184-023-00033-y>
- [4] (2025) Programme - dmhw 2025. SECI Congress. Accessed: 2025-08-13. [Online]. Available: <https://granada-en.congresoseci.com/dmhw2025/programme>
- [5] C. Potts, C. Kealy, J. M. McNulty, A. Madrid-Cagigal, T. Wilson, M. D. Mulvenna, S. O’Neill, G. Donohoe, and M. M. Barry, “Digital mental health interventions for young people aged 16-25 years: Scoping review,” *J Med Internet Res*, vol. 27, p. e72892, May 2025.
- [6] H. Li, R. Zhang, Y. C. Lee *et al.*, “Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being,” *npj Digital Medicine*, vol. 6, p. 236, 2023. [Online]. Available: <https://doi.org/10.1038/s41746-023-00979-5>
- [7] J. Alonso, M. Angermeyer, S. Bernert, R. Bruffaerts, T. Brugha, H. Bryson, G. de Girolamo, R. Graaf, K. Demyttenaere, I. Gasquet, J. Haro, S. Katz, R. Kessler, V. Kovess, J. Lépine, J. Ormel, G. Polidori, L. Russo, G. Vilagut, J. Almansa, S. Arbabzadeh-Bouchez, J. Autonell, M. Bernal, M. Buist-Bouwman, M. Codony, A. Domingo-Salvany, M. Ferrer, S. Joo, M. Martínez-Alonso, H. Matschinger, F. Mazzi, Z. Morgan, P. Morosini, C. Palacín, B. Romera, N. Taub, W. Vollebergh, and ESEMeD/MHEDEA 2000 Investigators, “Disability and quality of life impact of mental disorders in Europe: results

- from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project,” *Acta Psychiatrica Scandinavica Supplementum*, no. 420, pp. 38–46, 2004.
- [8] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, “Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication,” *Archives of General Psychiatry*, vol. 62, no. 6, pp. 593–602, Jun 2005, erratum in: *Arch Gen Psychiatry*. 2005 Jul;62(7):768.
- [9] J. Alonso, M. Angermeyer, S. Bernert, R. Bruffaerts, T. Brugha, H. Bryson, G. de Girolamo, R. Graaf, K. Demyttenaere, I. Gasquet, J. Haro, S. Katz, R. Kessler, V. Kovess, J. Lépine, J. Ormel, G. Polidori, L. Russo, G. Vilagut, J. Almansa, S. Arbabzadeh-Bouchez, J. Autonell, M. Bernal, M. Buist-Bouwman, M. Codony, A. Domingo-Salvany, M. Ferrer, S. Joo, M. Martínez-Alonso, H. Matschinger, F. Mazzi, Z. Morgan, P. Morosini, C. Palacín, B. Romera, N. Taub, W. Vollebergh, and ESEMeD/MHEDEA 2000 Investigators, “Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project,” *Acta Psychiatrica Scandinavica Supplementum*, no. 420, pp. 21–27, 2004.
- [10] Institute for Health Metrics and Evaluation, “Global burden of disease study,” <https://www.healthdata.org/gbd>, 2024, data source: IHME, Global Burden of Disease (2024).
- [11] Institute for Health Metrics and Evaluation (IHME), “Global burden of disease research & analysis,” <https://www.healthdata.org/research-analysis/gbd>, 2025, accessed 29 June 2025.
- [12] Health Knowledge, “Measures of disease burden,” <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/measures-disease-burden>, 2025, accessed 29 June 2025.
- [13] Statista, “Mental Health in Europe,” <https://www.statista.com/topics/7916/mental-health-in-europe/>, 2025, accessed 29 June 2025.
- [14] World Health Organization, “New WHO guidance calls for urgent transformation of mental health policies,” <https://www.who.int/news/item/25-03-2025-new-who-guidance-calls-for-urgent-transformation-of-mental-health-policies>, Mar. 2025, accessed 29 June 2025.
- [15] —, “Mental health,” <https://www.who.int/health-topics/mental-health>, 2025, accessed 9 April 2025.
- [16] —, “Schizophrenia,” <https://www.who.int/news-room/fact-sheets/detail/schizophrenia>, 2025, accessed 29 June 2025.

- [17] M. Birchwood, M. Michail, A. Meaden, N. Tarrier, S. Lewis, T. Wykes, L. Davies, G. Dunn, and E. Peters, “Cognitive behaviour therapy to prevent harmful compliance with command hallucinations (command): a randomised controlled trial,” *The Lancet Psychiatry*, vol. 1, no. 1, pp. 23–33, 2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2215036614702470>
- [18] E. O. Nwoye, A. A. Muslehat, C. Umeh, S. O. Okodeh, and W. L. Woo, “Schizobot: Delivering cognitive behavioural therapy for augmented management of schizophrenia,” *Digital Technologies Research and Applications*, vol. 3, no. 2, p. 24–40, Apr. 2024. [Online]. Available: <https://ojs.ukscip.com/journals/dtra/article/view/225>
- [19] R. Dale, “The return of the chatbots,” *Natural Language Engineering*, vol. 22, pp. 811–817, 09 2016.
- [20] A. Følstad and P. B. Brandtzæg, “Chatbots and the new world of hci,” *Interactions*, vol. 24, no. 4, p. 38–42, Jun. 2017. [Online]. Available: <https://doi.org/10.1145/3085558>
- [21] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed., 2025, online manuscript released August 24, 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
- [22] L. Laranjo, A. Dunn, H. L. Tong, A. Kocaballi, J. A. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Lau, and E. Coiera, “Conversational agents in healthcare: a systematic review,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 25, pp. 1248 – 1258, 2018.
- [23] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, “A survey of available corpora for building data-driven dialogue systems,” 2017. [Online]. Available: <https://arxiv.org/abs/1512.05742>
- [24] S. Young, M. Gašić, B. Thomson, and J. D. Williams, “Pomdp-based statistical spoken dialog systems: A review,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.
- [25] H. Gaffney, W. Mansell, and S. Tai, “Conversational agents in the treatment of mental health problems: mixed-method systematic review,” *JMIR mental health*, vol. 6, no. 10, p. e14166, 2019.
- [26] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, and H. Yan, “Designing embodied conversational agents,” *Embodied conversational agents*, vol. 29, 2000.

- [27] L. Zou, “Chapter 6 - meta-learning for natural language processing,” in *Meta-Learning*, L. Zou, Ed. Academic Press, 2023, pp. 209–266. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780323899314000055>
- [28] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 02 2008, vol. 2.
- [29] A. Ritter, C. Cherry, and W. B. Dolan, “Data-driven response generation in social media,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, R. Barzilay and M. Johnson, Eds. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 583–593. [Online]. Available: <https://aclanthology.org/D11-1054/>
- [30] H. Chen, X. Liu, D. Yin, and J. Tang, “A survey on dialogue systems: Recent advances and new frontiers,” *CoRR*, vol. abs/1711.01731, 2017. [Online]. Available: <http://arxiv.org/abs/1711.01731>
- [31] H. Brabra, M. Báez, B. Benatallah, W. Gaaloul, S. Bouguelia, and S. Zamanirad, “Dialogue management in conversational systems: A review of approaches, challenges, and opportunities,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 783–798, 2022.
- [32] J. Levis and R. Suvorov, “Automatic speech recognition,” in *The Encyclopedia of Applied Linguistics*, 2020, pp. 1–8. [Online]. Available: <https://doi.org/10.1002/9781405198431.WBEAL0066>
- [33] C.-M. Karat, C. Halverson, D. Horn, and J. Karat, “Patterns of entry and correction in large vocabulary continuous speech recognition systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’99. New York, NY, USA: Association for Computing Machinery, 1999, p. 568–575. [Online]. Available: <https://doi.org/10.1145/302979.303160>
- [34] M. K. Wolters, F. Kelly, and J. Kilgour, “Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia,” *Health Informatics Journal*, vol. 22, no. 4, pp. 854–866, 2016, PMID: 26276794. [Online]. Available: <https://doi.org/10.1177/1460458215593329>
- [35] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, “Conversational agents in healthcare: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 07 2018. [Online]. Available: <https://doi.org/10.1093/jamia/ocy072>

- [36] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial,” *JMIR Ment Health*, vol. 4, no. 2, p. e19, Jun 2017.
- [37] J. Weizenbaum, “Eliza—a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- [38] C. Stiles-Shields. (2019) Woebot: A professional review. Accessed: 2025-06-03. [Online]. Available: <https://onemindpsyberguide.org/guide/expert-review/woebot-an-expert-review/>
- [39] J. Hagen. (2025) Woebot health is shutting down its app. Accessed: 2025-06-03. [Online]. Available: <https://www.mobihealthnews.com/news/woebot-health-shutting-down-its-app>
- [40] M. Rauws. (2019) X2 executive summary. Accessed: 2025-06-03. [Online]. Available: <https://www.x2ai.com/blog/x2-executive-summary>
- [41] A. Joerin, M. Rauws, and M. L. Ackerman, “Psychological artificial intelligence service, tess: delivering on-demand support to patients and their caregivers: technical report,” *Cureus*, vol. 11, no. 1, 2019.
- [42] T. N. Stephens, A. Joerin, M. Rauws, and L. N. Werk, “Feasibility of pediatric obesity and prediabetes treatment support through tess, the ai behavioral coaching chatbot,” *Translational behavioral medicine*, vol. 9, no. 3, pp. 440–447, 2019.
- [43] R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, M. Rauws *et al.*, “Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial,” *JMIR mental health*, vol. 5, no. 4, p. e9782, 2018.
- [44] Wysa Ltd. (2025) Wysa – everyday mental health. Accessed: 2025-06-03. [Online]. Available: <https://www.wysa.com/>
- [45] B. Inkster, S. Sarda, and V. Subramanian, “An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study,” *JMIR Mhealth Uhealth*, vol. 6, no. 11, p. e12106, Nov 2018.
- [46] Business Wire. (2022) Wysa receives fda breakthrough device designation for ai-led mental health conversational agent. Accessed April 15, 2025.
- [47] S. Karkosz, R. Szymański, K. Sanna, and J. Michałowski, “Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: Randomized controlled trial,” *JMIR Form Res*, vol. 8, p. e47960, Mar 2024.

- [48] D. Gratzner. (2025) Reading of the week: Ai chatbots for mental healthcare – the new nejm ai paper; also, neuromodulation for schizophrenia (ajp). Accessed: 2025-06-03. [Online]. Available: <https://davidgratzner.com/reading-of-the-week/reading-of-the-week-ai-chatbots-for-mental-healthcare-the-new-nejm-ai-paper-also-neuromodulation-for-schizophrenia-ajp/>
- [49] M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, H. A. Banta, A. D. Jewett, A. J. Salzhauer, T. Z. Griffin, and N. C. Jacobson, “Randomized trial of a generative ai chatbot for mental health treatment,” *NEJM AI*, vol. 2, no. 4, p. AIoa2400802, 2025. [Online]. Available: <https://ai.nejm.org/doi/full/10.1056/AIoa2400802>
- [50] Z. Yi, J. Ouyang, Y. Liu, T. Liao, Z. Xu, and Y. Shen, “A survey on recent advances in LLM-based multi-turn dialogue systems,” *arXiv preprint arXiv:2402.18013*, 2024. [Online]. Available: <https://arxiv.org/abs/2402.18013>
- [51] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” 2020. [Online]. Available: <https://arxiv.org/abs/2001.08361>
- [52] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020. [Online]. Available: <https://doi.org/10.1007/s11431-020-1647-3>
- [53] S. Park, H. Subramonyam, and C. Kulkarni, “Thinking assistants: LLM-based conversational assistants that help users think by asking rather than answering,” *arXiv preprint arXiv:2312.06024*, 2024. [Online]. Available: <https://arxiv.org/abs/2312.06024>
- [54] L. M. Vowels, R. R. Francois-Walcott, and J. Darwiche, “Ai in relationship counselling: Evaluating chatGPT’s therapeutic capabilities in providing relationship advice,” *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100078, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2949882124000380>
- [55] F. Frisone, C. Pupillo, C. Rossi, and G. Riva, “Socrates. developing and evaluating a fine-tuned chatgpt model for accessible mental health intervention,” *Cyberpsychology, Behavior, and Social Networking*, vol. 0, no. 0, p. null, 0, pMID: 40241613. [Online]. Available: <https://doi.org/10.1089/cyber.2025.45510.cyeuro>
- [56] M. Luo, C. J. Warren, L. Cheng, H. M. Abdul-Muhsin, and I. Banerjee, “Assessing empathy in large language models with real-world physician-patient interactions,” *arXiv preprint arXiv:2405.16402*, 2024. [Online]. Available: <https://arxiv.org/abs/2405.16402>

- [57] A. Yuan, E. Garcia Colato, B. Pescosolido, H. Song, and S. Samtani, “Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots,” *ACM Trans. Manage. Inf. Syst.*, vol. 16, no. 1, Feb. 2025. [Online]. Available: <https://doi.org/10.1145/3701041>
- [58] The Brussels Times. (2023) Belgian man commits suicide following exchanges with chatgpt. Accessed: 2025-07-01. [Online]. Available: <https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt>
- [59] A. T. Duchowski, *Eye Tracking: Methodology, Theory and Practice*. Springer International Publishing AG, 2017.
- [60] A. Andrychowicz-Trojanowska, “Basic terminology of eye-tracking research,” *Applied Linguistics Papers*, vol. 25, no. 2, pp. 123–132, 2018, published by the University of Warsaw.
- [61] F. Mulvey *et al.*, “Exploration of safety issues in eye tracking,” 2008, cOGAIN EU Network of Excellence.
- [62] Bitbrain, “The different kinds of eye tracking devices,” June 2020, accessed May 6, 2025. [Online]. Available: <https://www.bitbrain.com/blog/eye-tracking-devices>
- [63] A. Bojko, *Eye Tracking the User Experience: A Practical Guide to Research*. Rosenfeld, 12 2013.
- [64] Tobii, “Eye tacker tobii pro x2-30,” <https://www.tobii.com/products/discontinued/tobii-pro-x2-30/>, accessed: 2025-08-18.
- [65] A. M. Mento and T. B. Team, “Different kinds of eye tracking devices,” <https://www.bitbrain.com/blog/eye-tracking-devices>, May 2025, accessed: 2025-08-18.
- [66] SR Research Ltd., “Eyelink 1000 plus – eye tracker,” <https://www.sr-research.com/eyelink-1000-plus/>, accessed: 2025-08-18.
- [67] Pupil Labs GmbH, “Eye tracking glasses – state of the art, affordable, accessible,” <https://pupil-labs.com/products>, 2025, accessed: 2025-08-18.
- [68] Eyeware. (2022) Eye tracking 101: What is it & how does it work in real life? Accessed: 2025-06-03. [Online]. Available: <https://eyeware.tech/blog/what-is-eye-tracking/>
- [69] Tobii, “Eye tracker sampling frequency,” [https://connect.tobii.com/s/article/eye-tracker-sampling-frequency?language=en\\_US](https://connect.tobii.com/s/article/eye-tracker-sampling-frequency?language=en_US), accessed: 2025-08-18.
- [70] M. Juhola, V. Jäntti, and I. Pyykkö, “Effect of sampling frequencies on computation of the maximum velocity of saccadic eye movements,” *Biological Cybernetics*, vol. 53, no. 2, pp. 67–72, Dec 1985. [Online]. Available: <https://doi.org/10.1007/BF00337023>

- [71] T. J. Crawford, S. Higham, T. Renvoize, J. Patel, M. Dale, A. Suriya, and S. Tetley, “Inhibitory control of saccadic eye movements and cognitive impairment in alzheimer’s disease,” *Biological Psychiatry*, vol. 57, no. 9, pp. 1052–1060, May 2005.
- [72] R. G. Alexander, S. L. Macknik, and S. Martinez-Conde, “Microsaccade characteristics in neurological and ophthalmic disease,” *Frontiers in neurology*, vol. 9, p. 144, 2018.
- [73] K. A. Dalrymple, M. D. Manner, K. A. Harmelink, E. P. Teska, and J. T. Ellison, “An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood,” *Frontiers in Psychology*, vol. Volume 9 - 2018, 2018. [Online]. Available: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.00803>
- [74] A. M. Feit, S. Williams, A. Toledo, A. Paradiso, H. Kulkarni, S. Kane, and M. R. Morris, “Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design,” in *Proceedings of the 2017 Chi conference on human factors in computing systems*, 2017, pp. 1118–1130.
- [75] M. Just and P. Carpenter, “Eye fixations and cognitive processes,” *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, 1976. [Online]. Available: [https://doi.org/10.1016/0010-0285\(76\)90015-3](https://doi.org/10.1016/0010-0285(76)90015-3)
- [76] R. Krueger, S. Koch, and T. Ertl, “Saccadelenses: interactive exploratory filtering of eye tracking trajectories,” in *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*, 2016, pp. 31–34.
- [77] Tobii AB. Different types of eye movements. Accessed: 2025-06-07. [Online]. Available: [https://connect.tobii.com/s/article/types-of-eye-movements?language=en\\_US](https://connect.tobii.com/s/article/types-of-eye-movements?language=en_US)
- [78] S. Martinez-Conde, S. Macknik, and D. Hubel, “The role of fixational eye movements in visual perception,” *Nature reviews. Neuroscience*, vol. 5, pp. 229–40, 04 2004.
- [79] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, “The role of fixational eye movements in visual perception,” *Nature Reviews Neuroscience*, vol. 5, pp. 229–240, Mar. 2004. [Online]. Available: <https://doi.org/10.1038/nrn1348>
- [80] M. Nyström, R. Andersson, D. C. Niehorster, R. S. Hessels, and I. T. C. Hooge, “What is a blink? classifying and characterizing blinks in eye openness signals,” *Behavior Research Methods*, vol. 56, no. 4, pp. 3280–3299, 2024. [Online]. Available: <https://doi.org/10.3758/s13428-023-02333-9>
- [81] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan, “Eye tracking and online search: Lessons learned and challenges

- ahead,” *Journal of the American Society for Information Science and Technology*, vol. 59, no. 7, pp. 1041–1052, 2008.
- [82] G. Bailly, F. Elisei, S. Raidt, A. Casari, and A. Picot, “Embodied conversational agents: computing and rendering realistic gaze patterns,” in *Pacific-Rim Conference on Multimedia*. Springer, 2006, pp. 9–18.
- [83] N. Bee, J. Wagner, E. André, T. Vogt, F. Charles, D. Pizzi, and M. Cavazza, “Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application,” in *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, 2010, pp. 1–8.
- [84] K. R. Thórisson, “Gandalf: An embodied humanoid capable of real-time multimodal dialogue with people,” in *Agents*, 1997, pp. 536–537.
- [85] T. Amorese, C. Greco, M. Cuciniello, C. Buono, C. Palmero, P. Buch-Cardona, S. Escalera, M. I. Torres, G. Cordasco, and A. Esposito, “Using eye tracking to investigate interaction between humans and virtual agents,” in *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 2022, pp. 125–132.
- [86] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, “The pupil as a measure of emotional arousal and autonomic activation,” *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.2008.00654.x>
- [87] G. Lee, A. Ojha, and M. Lee, “Concentration monitoring for intelligent tutoring system based on pupil and eye-blink,” in *Proceedings of the 3rd International Conference on Human-Agent Interaction*, ser. HAI ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 291–294. [Online]. Available: <https://doi.org/10.1145/2814940.2815000>
- [88] R. Vertegaal, “Attentive user interfaces,” *Communications of the ACM*, vol. 46, 03 2003.
- [89] H. Prendinger, A. Hyrskykari, M. Nakayama *et al.*, “Attentive interfaces for users with disabilities: eye gaze for intention and uncertainty estimation,” *Univ Access in the Inf Soc*, vol. 8, pp. 339–354, 2009. [Online]. Available: <https://doi.org/10.1007/s10209-009-0144-5>
- [90] J. S. Shell, T. Selker, and R. Vertegaal, “Interacting with groups of computers,” *Communications of the ACM*, vol. 46, no. 3, pp. 40–46, 2003.
- [91] R. Vertegaal, J. S. Shell, D. Chen, and A. Mamuji, “Designing for augmented attention: Towards a framework for attentive user interfaces,” *Computers in Human Behavior*,

- vol. 22, no. 4, pp. 771–789, 2006, attention aware systems. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0747563205001160>
- [92] S. Duncan, “Some signals and rules for taking speaking turns in conversations,” *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.
- [93] A. Hyrskykari, P. Majaranta, and K.-J. Räihä, “From gaze control to attentive interfaces,” in *Proceedings of HCII*, vol. 2. Citeseer, 2005, p. 9.
- [94] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol, “DIET: Lightweight language understanding for dialogue systems,” 2020. [Online]. Available: <https://arxiv.org/abs/2004.09936>
- [95] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [96] A. Zygadło, M. Kozłowski, and A. Janicki, “Text-based emotion recognition in english and Polish for therapeutic chatbot,” *Applied Sciences (Switzerland)*, vol. 11, no. 21, 2021.
- [97] V. Vlasov, J. E. M. Mosig, and A. Nichol, “Dialogue transformers,” 2019. [Online]. Available: <https://arxiv.org/abs/1910.00486>
- [98] A. Jiao, “An intelligent chatbot system based on entity extraction using RASA NLU and neural network,” *Journal of Physics: Conference Series*, vol. 1487, no. 1, p. 012014, mar 2020.
- [99] World Health Organization. (2025) Icd-10 version:2019 - f20.0 paranoid schizophrenia. Accessed on 2025-07-05. [Online]. Available: <https://icd.who.int/browse10/2019/en#F20.0>
- [100] M. J. Rantala, S. Luoto, J. I. Borráz-León, and I. Krams, “Schizophrenia: The new etiological synthesis,” *Neuroscience & Biobehavioral Reviews*, vol. 142, p. 104894, 2022.
- [101] M. Hany and A. Rizvi. (2025) Schizophrenia. Treasure Island (FL). [Updated 2024 Feb 23]. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK539864/>
- [102] D. M. Barch and H. Berenbaum, “Language generation in schizophrenia and mania: the relationships among verbosity, syntactic complexity, and pausing,” *Journal of Psycholinguistic Research*, vol. 26, no. 4, pp. 401–412, Jul. 1997.
- [103] H. Karbe, “Wernicke’s Area,” in *Encyclopedia of the Neurological Sciences*. Elsevier Inc., jan 2014, pp. 751–752.

- [104] K. Harezlak, P. Kasproski, and M. Stasch, “Towards accurate eye tracker calibration – methods and procedures,” *Procedia Computer Science*, vol. 35, pp. 1073–1081, 2014, knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050914011594>
- [105] P. Kasproski and K. Harezlak, “Using the uncalibrated eye tracker signal in the gaming environment,” in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA ’20 Short Papers. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: <https://doi.org/10.1145/3379156.3391837>
- [106] J. Beatty and B. Lucero-Wagoner, “The pupillary system,” in *Handbook of Psychophysiology*, 2nd ed., J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds. Cambridge: Cambridge University Press, 2000, pp. 142–162.
- [107] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, “Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze,” *PLOS ONE*, vol. 13, no. 9, pp. 1–23, 09 2018. [Online]. Available: <https://doi.org/10.1371/journal.pone.0203629>
- [108] B. Laeng, S. Sirois, and G. Gredebäck, “Pupillometry: A window to the preconscious?” *Perspectives on Psychological Science*, vol. 7, no. 1, pp. 18–27, 2012.
- [109] J. J. Castro, A. M. Pozo, M. Rubiño, R. G. Anera, and L. Jiménez Del Barco, “Retinal-image quality and night-vision performance after alcohol consumption,” *Journal of Ophthalmology*, vol. 2014, p. 704823, 2014.
- [110] M. Oliva and A. Anikin, “Pupil dilation reflects the time course of emotion recognition in human vocalizations,” *Scientific Reports*, vol. 8, p. 4871, 2018.
- [111] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, “The pupil as a measure of emotional arousal and autonomic activation,” *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.
- [112] H. Mitre-Hernandez, R. Covarrubias Carrillo, and C. Lara-Alvarez, “Pupillary responses for cognitive load measurement to classify difficulty levels in an educational video game: Empirical study,” *JMIR Serious Games*, vol. 9, no. 1, p. e21620, Jan 2021.
- [113] A. P. Belliveau, A. N. Somani, and R. H. Dossani, “Pupillary light reflex,” in *StatPearls [Internet]*. StatPearls Publishing, 2023.
- [114] P. Schober and T. R. Vetter, “Confidence intervals in clinical research,” *Anesthesia & Analgesia*, vol. 130, no. 5, pp. 1303–1304, May 2020.

- [115] K. Gabor-Siatkowska and A. Janicki, “Czy kalibracja okulografu w interfejsie człowiek-komputer zawsze jest niezbędna?” *Przegląd Telekomunikacyjny+ Wiadomości Telekomunikacyjne*, 2022.
- [116] —, “Badanie zasadności używania kalibracji w pomiarach okulograficznych w celu zastosowania w interfejsie człowiek-komputer,” *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*, 2023.
- [117] A. Kendon, “Some functions of gaze-direction in social interaction,” *Acta Psychologica*, vol. 26, pp. 22–63, 1967, [Online; accessed July 6, 2025]. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0001691867900054>
- [118] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, “Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 301–308. [Online]. Available: <https://doi.org/10.1145/365024.365119>
- [119] C. Breil, L. Huestegge, and A. Böckler, “From eye to arrow: Attention capture by direct gaze requires more than just the eyes,” *Attention, Perception, & Psychophysics*, vol. 84, pp. 64–75, 2022. [Online]. Available: <https://doi.org/10.3758/s13414-021-02382-2>
- [120] J. Arizpe, V. Walsh, G. Yovel, and C. I. Baker, “The categories, frequencies, and stability of idiosyncratic eye-movement patterns to faces,” *Vision Research*, vol. 141, pp. 191–203, 2017. [Online]. Available: <https://doi.org/10.1016/j.visres.2016.10.013>
- [121] K. Gabor-Siatkowska, I. Stefaniak, and A. Janicki, “A multimodal approach for improving a dialogue agent for therapeutic sessions in psychiatry,” in *Transforming Media Accessibility in Europe*. Springer, 2024, pp. 397–414.
- [122] —, “Gaze-dependent response activation in dialogue agent for cognitive-behavioral therapy,” *Procedia Computer Science*, vol. 246, pp. 2322–2331, 2024.
- [123] R. Howes. (2012) Eye contact in therapy, part i. Psychology Today. Accessed: 2025-07-20. [Online]. Available: <https://www.psychologytoday.com/us/blog/in-therapy/201202/eye-contact-in-therapy-part->
- [124] OpenAI, “What’s new in assistants api beta,” <https://platform.openai.com/docs/assistants/whats-new>, accessed: 2025-08-22.
- [125] M. de la Vega, “Understanding openai’s “temperature” and “ $top_p$ ” parameters in language models,” <https://medium.com/@1511425435311/understanding-openais-temperature-and-top-p-parameters-in-language-models-d2066504684f>, Nov. 2023, accessed: 2025-08-22.

- [126] O. Community. (2023, Apr.) Cheat sheet: Mastering temperature and *top\_p* in ChatGPT api. [Online]. Available: <https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api/172683>
- [127] IBM. (2024, Oct.) What is rag (retrieval-augmented generation)? <https://www.ibm.com/think/topics/retrieval-augmented-generation>.
- [128] D. L. Klamen and R. Yudkowsky, “Using standardized patients for formative feedback in an introduction to psychotherapy course,” *Academic Psychiatry*, vol. 26, no. 3, pp. 168–172, 2002. [Online]. Available: <https://doi.org/10.1176/appi.ap.26.3.168>
- [129] J. A. Cleland, K. Abe, and J.-J. R. and, “The use of simulated patients in medical education: A mee guide no 42,” *Medical Teacher*, vol. 31, no. 6, pp. 477–486, 2009, PMID: 19811162. [Online]. Available: <https://doi.org/10.1080/01421590903002821>
- [130] Teatr Politechniki Warszawskiej. (2025) Teatr politechniki warszawskiej. [Online]. Available: <https://www.teatr.pw.edu.pl/>
- [131] R. Masumura, T. Tanaka, A. Ando, R. Ishii, R. Higashinaka, and Y. Aono, “Neural dialogue context online end-of-turn detection,” in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, K. Komatani, D. Litman, K. Yu, A. Papangelis, L. Cavedon, and M. Nakano, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 224–228. [Online]. Available: <https://aclanthology.org/W18-5024/>
- [132] A. Raux and M. Eskenazi, “A finite-state turn-taking model for spoken dialog systems,” in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 629–637.
- [133] H. Khouzaimi, R. Laroche, and F. Lefevre, “Turn-taking phenomena in incremental dialogue systems,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1890–1895.
- [134] K. Gabor-Siatkowska, M. Sowański, M. Pudo, R. Rzatkiwicz, I. Stefaniak, M. Kozłowski, and A. Janicki, “Therapeutic spoken dialogue system in clinical settings: Initial experiments,” in *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2023, pp. 1–5.
- [135] M. Kozłowski, K. Gabor-Siatkowska, I. Stefaniak, M. Sowański, and A. Janicki, “Enhanced emotion and sentiment recognition for empathetic dialogue system using big data and deep learning methods,” in *International Conference on Computational Science*. Springer, 2023, pp. 465–480.

- [136] K. Gabor-Siatkowska, I. Stefaniak, and A. Janicki, “Beyond words: Gaze-enhanced IIm-based dialogue system for therapeutic purposes,” in *Proceedings of the 32nd International Conference on Systems, Signals and Image Processing (IWSSIP 2025)*. Skopje, North Macedonia: IEEE, 2025, pp. 1–5, (in press).
- [137] D. Chen and R. Vertegaal, “Using mental load for managing interruptions in physiologically attentive user interfaces,” in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 1513–1516. [Online]. Available: <https://doi.org/10.1145/985921.986103>
- [138] P. Kasprowski and K. Harezlak, “Using non-calibrated eye movement data to enhance human computer interfaces,” in *Intelligent Decision Technologies*, R. Neves-Silva, L. C. Jain, and R. J. Howlett, Eds. Cham: Springer International Publishing, 2015, pp. 347–356.
- [139] K. Harezlak, A. Duliban, and P. Kasprowski, “Eye movement-based methods for human-system interaction. a comparison of different approaches,” *Procedia Computer Science*, vol. 192, pp. 3099–3108, 2021, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050921018196>
- [140] M. Duszczyk. (2025, aug) Ai może doprowadzić do groźnej choroby. eksperci ostrzegają. Rzeczpospolita (Cyfrowa). [Online]. Available: <https://cyfrowa.rp.pl/technologie/art42835751-ai-moze-doprowadzic-do-groznej-choroby-eksperci-ostrzegaja>
- [141] OpenAI. (2025, aug) Jak zapewnić pomoc, kiedy jest najbardziej potrzebna. Polska wersja artykułu „Helping people when they need it most” opublikowana przez OpenAI. [Online]. Available: <https://openai.com/pl-PL/index/helping-people-when-they-need-it-most/>