Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 111-138
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Autonomous VR Exposure Therapy for Anxiety Disorders Using Conversational AI—Ethical and Technical Implications and Preliminary Results

David Obremski, Carolin Wienrich

Psychology of Intelligent Interactive Systems, University of Würzburg, Würzburg, Germany
Email: david.obremski@uni-wuerzburg.de

How to cite this paper: Obremski, D. and Wienrich, C. (2024) Autonomous VR Exposure Therapy for Anxiety Disorders Using Conversational AI—Ethical and Technical Implications and Preliminary Results. *Journal of Computer and Communications*, **12**, 111-138.
https://doi.org/10.4236/jcc.2024.124010

Received: March 1, 2024
Accepted: April 20, 2024
Published: April 23, 2024

Abstract

The need for psychotherapy is very high and the lack of care causes a lot of suffering and high costs. This paper presents an interdisciplinary approach to creating an AI-guided exposure therapy for fear of heights in virtual reality (VR). First, ethical principles for the use of conversational AI in psychotherapy were translated into technical requirements and made measurable. Based on this, an autonomous virtual reality exposure therapy was iteratively developed with a therapist. The feasibility and implementation of the ethical principles were tested with a patient. The patient was very satisfied with the VR setup. The AI therapist was also rated positively, although there is still room for improvement regarding conversational skills. Overall, the paper shows how AI can contribute responsibly to improving the psycho-therapeutic supply. It also provides guidelines that make ethical principles tangible and measurable for developers.

Keywords

AI Therapy, Psychotherapy, Virtual Reality Exposure Therapy, Cognitive Behavioral Therapy

1. Introduction

With a 12-month prevalence of 14%, anxiety disorders are the mental illnesses with the highest prevalence in Europe [1]. Psychotherapy, especially exposure therapy, is considered an effective form of treatment alongside medication [2]. In contrast to the high demand for therapy places, there are also bottlenecks in the supply in Germany leading to an average waiting time of around five months to start therapy [3]. These waiting times are not only stressful for patients, but also generate enormous economic costs. In the EU alone, costs of 70 billion euros are incurred annually [4] [5]. While politics is helping to improve the situation, more and more technical and digital innovations are available to support the supply. Above all, Virtual Reality (VR) offers a controllable, safe and effective environment for exposure therapy [2]. Numerous studies prove its effectiveness and show cost savings (see [6]). This relieves the therapists, for whom exposure therapy, which often has to be carried out outside the medical office, is very time-consuming [7]. Beyond that, exposure therapies in VR receive a high level of acceptance among patients [2]. While the transfer of exposure therapy to VR leads to considerable time savings and thus helps to relieve the system, therapeutic supervision is still essential for its implementation. The therapist must conduct the pre- and post-talk, as well as guide the exposure and react to the patient and adapt the exposure accordingly. This does not yet eliminate the bottleneck and the potential of technology-supported therapy has not yet been fully exploited. VR alone cannot offer fully automated therapy components to relieve the shortage of therapists.

The recent rapid progress in the field of Artificial Intelligence (AI) in general and conversational AI in particular paves the way for increasing automation of exposure therapy components. With these systems being able to process large contexts of user data (e.g., GPT 4[1]) and to follow specific instructions on how to react to different user inputs, they might have the potential to guide patients through an exposure exercise in the form of embodied AI.

Before, at least partly, delegating the sensitive task of conducting therapy sessions with anxiety patients to a conversational AI, many ethical aspects and therapeutic standards have to be considered [8]. Additionally, these ethical aspects and therapeutic standards have to be translated to technical requirements and prompting instructions, to create guidelines that can be adhered to for developing safe and ethically acceptable AI in therapy settings.

Contribution: This paper first provides an overview of relevant work in the field of Virtual Reality Exposure Therapy (VRET) and presents existing AI-based approaches in therapy-related settings. In addition, it presents current ethical positions on the use of autonomous embodied AI in therapy contexts and derives technical requirements from them. Therapeutic standards have been incorporated into prompting instructions. Finally, the implementation and subsequent preliminary evaluation of a prototype realizing VRET with an AI-controlled agent as a virtual therapist is presented. This prototype was implemented iteratively with a therapist to test the general practicability of such a system, adhering to the previously defined guidelines, and extending the theoretically derived requirements based on experiences from a practical evaluation with a patient.

2. Related Work

According to the German chamber of psychotherapists, every third person suf-

[1] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo.

fers from a psychological disorder throughout their life, regardless of age, gender or cultural background [9]. In addition to the suffering of those affected, over 70 billion euros in costs are incurred annually across the EU due to lost working hours and treatment costs, which also makes mental illness an economic burden. In Germany alone, the sick days taken due to mental illness increased by 404% for women and 377% for men from 1997 to 2022 [10]. *Anxiety disorders* have been identified as one of the most common mental disorders, affecting roughly four percent of the global population and around 25% of the population in Germany at least once in their lifetime [11] [12] [13]. Whilst fear and anxiety are normal and valuable reactions as they can keep us from dangerous or hurtful situations, they turn into a *disorder* once they affect people's everyday life and prevent them from going about everyday activities, e.g., going grocery shopping, driving, or going to work. Anxiety disorders often lead to an avoidance of those scenarios, which is often accompanied with enormous suffering and restrictions [14].

Psychotherapy has been shown to be a strongly effective treatment approach for psychological disorders in general and anxiety disorders more specifically [15]. Within psychotherapy, a special form of cognitive behavior therapy, *exposure therapy*, has proven to be highly effective in treating anxiety disorders. It desensitizes patients by exposing them to the stimulus they are actively avoiding in a safe environment. During the exposition, patients learn that the expected amount of fear is higher than the actual fear they experience when being exposed to the respective stimulus. By learning that the stimulus (e.g., standing on a high bridge) doesn't lead to the result that is congruent with their expectations (e.g., falling from that bridge), their fear of this stimulus decreases [2] [16]. The exposure is embedded in exercises on perception, control and contextualization of anxiety. Exposure therapy is used as a focal practice within the treatment for mental disorders and has shown significant effects on affected patients [17].

While psychotherapy has shown to be a very effective way of treating mental disorders, the availability of therapy places is not in line with the demand. In Germany, patients wait an average of approximately 5 months for the commencement of psychotherapy after its necessity has been determined [3]. With the beginning of the COVID-19 pandemic the discrepancy between the demand of people seeking therapy and the shortage of therapists and treatment spots available has only grown [4] [5]. In an attempt to account for this lack of treatment opportunities, different strategies have been applied, from optimizing the system from a political perspective [18], to reforming the education to become a therapist to get more attractive[2]. In parallel to these efforts, new technology is increasingly deployed in therapeutic contexts.

Especially the medium VR in therapy contexts has been widely researched in the past decades [19] [20]. Although *in vivo* exposure therapy (*i.e.* experiencing the trigger of fear in real life) has proven more effective than imagined exposure,

[2] https://www.bundesgesundheitsministerium.de/psychotherapeutenausbildung.

various scenarios, e.g., flying, getting shot in the line of duty or being at a specific height, are difficult to recreate during a therapy session. Whilst most patients might benefit from being exposed to a stimuli that triggers their anxiety, the realistic recreation of some scenarios is either too time consuming, too expensive, or too dangerous. The term VRET describes the immersion of users in a VR setting, together with the controlled *exposure* to a stimulus previously avoided due to a prevailing psychological disorder, comparable to an exposure to a realistic stimulus [21]. VRET provides an affordable therapy tool to support psychotherapists within their treatments, specifically for exposure therapy [6] [22] [23]. VRET is highly accepted and sometimes even preferred by patients [24], while leading to similar results [6] [22] [23], compared to the rather costly in vivo therapy.

2.1. Virtual Reality Exposure Therapy for Anxiety Disorders

Various studies approached the development of virtual reality systems that are applicable within the psychological treatment of anxiety disorders.

In an attempt to validate the use of VR as a treatment method of anxiety disorders, various meta-reviews were conducted and addressed the effectiveness of VRET [6] [23] [25]. In 2008, Parsons and Rizzo examined 21 studies, with a total of 300 participants reporting anxiety levels pre and post-VRET measures [6]. They were able to show an overall large effect size over all studies (average random effect size for anxiety in total = 0.95) with lower anxiety values post-VRET, suggesting the significant effectiveness of VRET for anxiety disorders.

Powers and Emmelkamp also published a meta-review in the same year, including 13 studies with a total of 397 participants, which included the comparison of virtual reality exposure therapy with no treatment (*i.e.*, patients on the waiting list) and in vivo therapy (*i.e.* non-technology-mediated exposure therapy) [25]. They were able to show a large effect size ($d = 1.11$) for VRET in various domains (e.g., acrophobia, anxiety disorders) compared to no treatment and furthermore, even recognized a small effect ($d = 0.35$) of VR being more effective compared to non-technology-mediated therapy. In 2015, Morina *et al.* conducted a meta-review specifically focused on whether effects of VRET are generalizable to real-life. Their qualitative synthesis included 14 studies, which showed improvement of behaviour after VRET treatment compared to control groups undergoing no treatments [23]. Comparing treatments in VR to non-technology-mediated behavioral therapy, indicated no significant differences, revealing VRET to be an effective alternative to classic behavioral therapy.

The positive results of the meta analysis can also be seen in numerous individual studies [26] [27] [28]. In summary, it can be said that VRET is already an effective alternative to exposure in vivo for the treatment of anxiety disorders.

Nevertheless, VRET alone cannot eliminate the bottleneck of psychotherapy supply as a therapist is continuously involved. Each individual exercise is accompanied by a therapist limiting the potential of technology-supported therapy. Therefore, this paper looks at the additional potential of AI and shows in the following how subcomponents of the exposure therapy can be automated with an AI assistant and which technical requirements can be derived from ethical guidelines.

2.2. AI in Psychotherapy

The implementation of AI in psychotherapy can in principle be realized in two ways. It can be utilized in "in the loop" therapy, where the AI can trigger interventions defined, modified, or canceled by the therapist, or in "out-of-the-loop" therapy, without the involvement of human therapists in the interaction with the patient [29]. While the first approach uses AI more as a tool to improve psychotherapy, the second approach defines the role of AI as an independent agent [30]. With the rapid development of conversational AI in recent years [31], out-of-the-loop therapy has gone from a theoretical concept whose ethical implications have been discussed in various publications (e.g., [8] [32] [33]) to a topic that is actually technically feasible.

Utilizing the progress made in conversational AI, the first AI-based chatbots were implemented to support users' mental health [34] [35]. In 2018, the Lifeline project launched an AI-based chatbot for people struggling with suicide or self harm[3]. Targeting depression and anxiety, Woebot Health[4] offers a chatbot implementing practices of cognitive behavioral therapy to support people experiencing these symptoms. Fitzpatrick *et al.* conducted a randomized controlled trial comparing the effect of a two-week usage of the Woebot to that of reading a book about depression on students' symptoms of depression. The results show that the group using the chatbot showed a significant decrease of depression symptoms after the study, compared to the "information-only" group [35]. The chatbot Wysa[5] provides different chatbot solutions for both, adults and adolescents, to help them deal with topics such as anxiety, depression, or chronic pain. Evaluating the user feedback of people that have used the Wysa chatbot, Malik *et al.* found a high acceptance rate of the users, who especially valued its non-judgmentality and ease of conversation [36]. When assessing the therapeutic alliance of $N = 1205$ users, Beatty *et al.* found an increase of such during the usage of the chatbot, comparable with results of research on therapeutic alliance reported in the field of human-controlled psychotherapy [37].

While the use of AI in chatbots that target the improvement of the users' mental health is showing increasing popularity, the use of embodied AI to take on the role of a virtual therapist in VRET has not yet been investigated. The closest attempt to this was implemented by Freeman *et al.* in 2018. The authors investigated if an automated VRET controlled by a virtual avatar can reduce patients' fear of height [38]. While they report a positive effect of the intervention

[3] https://blog.twitter.com/en_au/topics/company/2018/Lifeline-launches-Twitter-DM-chatbot-to-help-BeALifeline.

[4] https://woebothealth.com/.

[5] https://www.wysa.com/.

on the users' fear of height, they used a pre-programmed scenario for the therapy session, rather than a dynamic one. Relying on pre-captured motion and speech of an actor that was applied to the virtual therapist afterwards, the users had very little possibility to influence the automated VRET. However, according to previous research, the individual relationship between therapist and patient is crucial for the outcome of psychotherapy [39], suggesting that scripted behaviour of the virtual avatar might have negative effects on the effect of the automated VRET.

With the current capabilities of conversational AI it could be possible to replace the generic virtual avatar by a fully autonomous virtual agent that, based on large language models, can interact with the patient in real time, establishing an emotional connection similar to the one observable between human therapists and their patients. Based on this highly flexible setting, the AI-therapist could also detect possible risks during the intervention and act accordingly to prevent harm.

Ultimately, the integration of AI into psycho-therapeutic services holds the potential to address service gaps, offering patients a low-threshold, consistently available access to psychotherapy [8] [32] [40]. This approach could reach individuals who might otherwise be unable to access therapy due to factors such as immobility or a low density of therapists [41]. It also includes individuals who reject psychotherapy due to fear of stigma or shame, as people tend to feel less shame towards non-anthropomorphic robots, leading to greater openness and a higher likelihood of utilizing their services [42] [43] [44] [45].

Besides the enormous possibilities, there are also considerable risks. Both are currently the subject of much discussion in various fields of application of generative AI, including psychotherapy [8] [40]. Numerous ethical principles are discussed here, such as beneficence, non-harm, or justice. However, it is difficult for developers to relate these ethical discussions to their specific implementation and use case and to understand exactly how they have to implement them technically. Thus, there is a lack of a translation of the ethical considerations into concrete technical requirements that enable safe and responsible development. To bridge this gap, we proceed to analyze ethical considerations to translate the associated risk-benefit assessment into manageable technical requirements guiding specific implementations and use cases.

3. Translation of Ethical into Technical Requirements

The application of AI in psychotherapy gives rise to crucial requirements for the respective systems and quality assurance. From the perspective of the German Ethics Council [40], it is generally imperative, when employing AI in the medical sector, to closely collaborate with medical professional societies and regulatory authorities throughout the system's development to deployment stages. This collaboration aims to identify and proactively address errors, minimizing or avoiding them altogether. Simultaneously, the system should undergo rigorous evaluation through appropriate testing, certification, and auditing measures, ensuring the fulfillment of minimum standards concerning autonomy, control, fairness, transparency, reliability, security, and data protection. Other experts in the field of psychotherapy, designate similarly the medical and socially ethical principles of respect for self-determination, beneficence, non-harm, justice, and epistemic (in)justice [8] [33] [46]. We have examined the overlaps between these principles and summarized them in five main categories: *beneficence*, *risk avoidance*, *autonomy*, *fairness*, and *involvement*. The principles are outlined below including corresponding potentials and risk and followed by the derivation of specific requirements for technical implementation (Figure 1). Please note that specific therapeutic standards have been described and incorporated below in the implementation of the promptings.

3.1. Beneficence

Essentially, the beneficience of an AI system is determined by its ability to exude trust worthiness and to be unbiased and impartial.

Trust and alliance: A central advantage, as well as a challenge, in employing AI in therapy is that psychotherapist supervision is not required for the treatment unit. Besides the treatment itself, the effectiveness of psychotherapy relies on

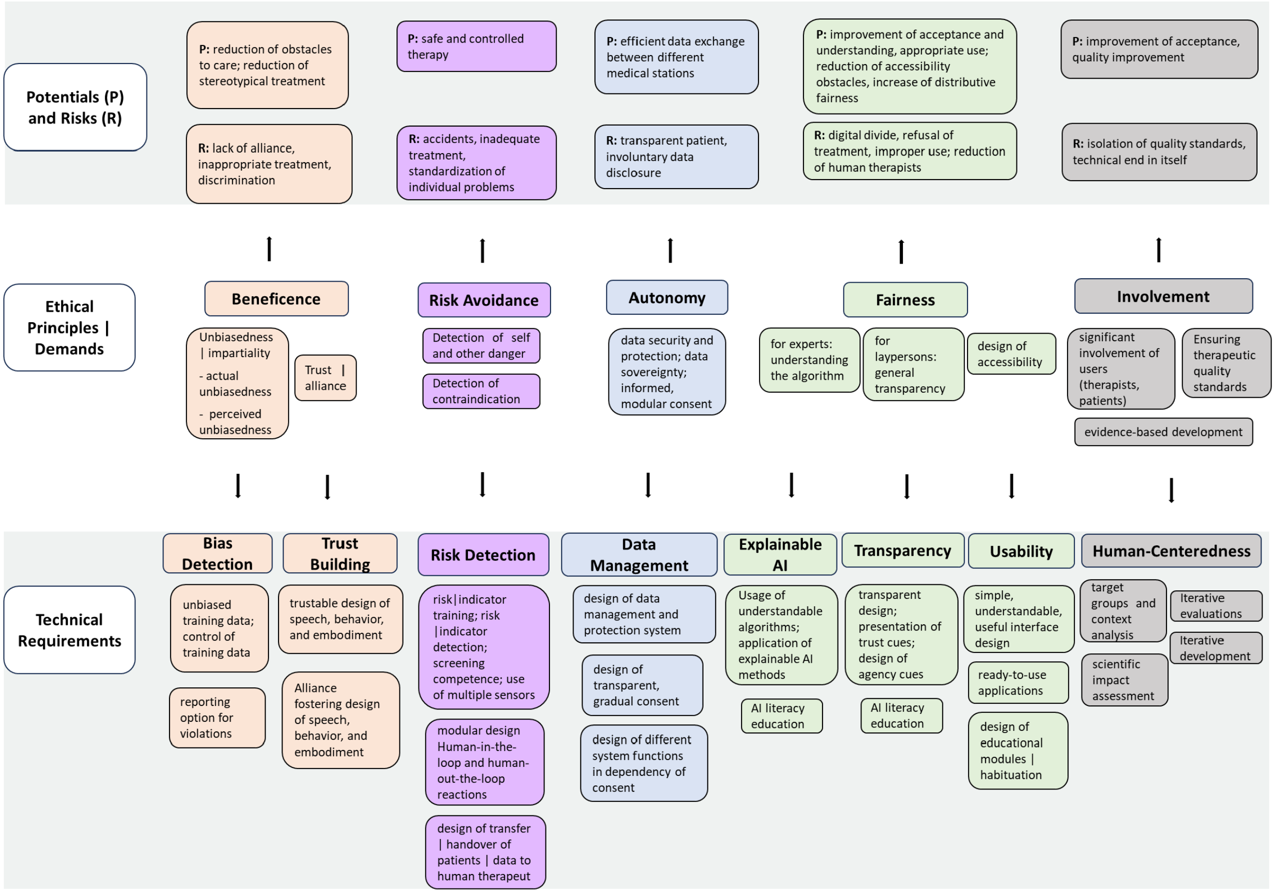
Potentials (P) and Risks (R)
P: reduction of obstacles to care; reduction of stereotypical treatment
R: lack of alliance, inappropriate treatment, discrimination
P: safe and controlled therapy
R: accidents, inadequate treatment, standardization of individual problems
P: efficient data exchange between different medical stations
R: transparent patient, involuntary data disclosure
P: improvement of acceptance and understanding, appropriate use; reduction of accessibility obstacles, increase of distributive fairness
R: digital divide, refusal of treatment, improper use; reduction of human therapists
P: improvement of acceptance, quality improvement
R: isolation of quality standards, technical end in itself
Ethical Principles | Demands
Beneficence
Unbiasedness | impartiality
- actual unbiasedness
- perceived unbiasedness
Trust | alliance
Risk Avoidance
Detection of self and other danger
Detection of contraindication
Autonomy
data security and protection; data sovereignty; informed, modular consent
Fairness
for experts: understanding the algorithm
for laypersons: general transparency
design of accessibility
Involvement
significant involvement of users (therapists, patients)
Ensuring therapeutic quality standards
evidence-based development
Technical Requirements
Bias Detection
unbiased training data; control of training data
reporting option for violations
Trust Building
trustable design of speech, behavior, and embodiment
Alliance fostering design of speech, behavior, and embodiment
Risk Detection
risk|indicator training; risk |indicator detection; screening competence; use of multiple sensors
modular design Human-in-the-loop and human-out-the-loop reactions
design of transfer | handover of patients | data to human therapeut
Data Management
design of data management and protection system
design of transparent, gradual consent
design of different system functions in dependency of consent
Explainable AI
Usage of understandable algorithms; application of explainable AI methods
AI literacy education
Transparency
transparent design; presentation of trust cues; design of agency cues
AI literacy education
Usability
simple, understandable, useful interface design
ready-to-use applications
design of educational modules | habituation
Human-Centeredness
target groups and context analysis
scientific impact assessment
Iterative evaluations
Iterative development

Figure 1. Overview of ethical principles, corresponding potentials, risks, and the derived requirements for system development.

the alliance between the patient and the therapist [39]. The strength of the alliance between therapist and patient positively correlates with therapy outcomes [47]. Trust, empathy, and appreciation are facilitating factors in alliance formation [47]. The inclusion of empathetic and prosocial behavior in a human-computer relationship can foster an alliance between humans and computers, suggesting that such alliances are not limited to interpersonal relationships [48]. Numerous studies confirm that people can form social relationships with technical and artificial entities, triggering similar social psychological mechanisms as in human-human interaction [49] [50].

To establish an emotional connection, build trust, and avoid negative feelings toward the AI, an emphasis should be placed on the embodiment of the AI and the personality it portrays. Female appearance of e.g., virtual medical agents has shown to be high in acceptance and trust evoking within participants [51]. When implementing speech of the embodied AI, previous research has suggested the impact of interaction effects on gender, showing more positive effects for trust if the gendered voice of the virtual agent was congruent with the participant's gender [52] [53]. Additionally, naturalness of speech of an IVA has shown to have significant effects on warmth [54]. Further, non-verbal behavior, such as gestures and facial expressions of the embodied AI can contribute to the embodied AI's perceived trustworthiness, as e.g., embodied virtual agents were perceived as more trustworthy when they were smiling [55]. Similar to human-human-interaction, the use of in group cues and perceived similarity with an IVA can elicit higher empathy towards embodied AI and therefore possibly improve interpersonal connection with the AI [56].

In summary, it can be seen that the alliance formation can be influenced by the language, appearance and behavior of the AI assistant, very similar to the human-human alliance.

Impartiality: Another important aspect of the principle of *beneficence* is the question of how strongly the algorithm is biased and discriminates against certain users. While AI is perceived as less stigmatizing, this quality is rooted in the data used for training. Thus, already in the development phase, particular attention must be given to the data used for feeding the algorithm [40]. Especially in the psycho-therapeutic context, developers must be aware of implicit "human biases" and "epistemic injustice" and control for them [8].

Weighing up potentials and risks: The implementation of *beneficence* is crucial to ensure a successful therapy by establishing a base of trust and understanding and forming an alliance between the patient and therapist. It can also reduce the risk of stereotypical treatments. If the ethical principle is not adhered to, this can have a negative impact on the outcome of therapy and lead to patients feeling misunderstood, unaccepted or discriminated.

Technical Requirements:

- **Trust building:** While it is important for the AI to act professionally, there also need to be preceding mechanisms in place that allow for the patient to form a relationship with the AI. The AI's appearance, verbal and non-verbal behavior should also promote the formation of trust.

- **Bias detection:** To ensure the use of unbiased data, methods such as "Bias Detectives" should be used to review the training data [8]. Additionally, the model should be specifically instructed to avoid biased behavior based on characteristics such as culture or gender.

3.2. Risk Avoidance

Psychotherapists bear the responsibility of not only addressing mental health concerns but also diligently recognizing and mitigating potential risks that may arise during the therapeutic process. They have the ethical responsibility to recognize and prevent risks for the patient and involved parties (self-endangerment and endangerment of others). Therefore, detecting suicidality, recognizing contraindications, managing symptom deterioration, or addressing the absence of necessary patient skills for therapy (e.g., self-reflection, frustration tolerance) are challenges that must be considered in development [8]. The detection of potential risks could be enhanced by sensors. When recognizing risks, an appropriate response from the AI must take place, potentially involving a warning function [40] or the implementation of warning stages and stop signals to allow the possibility to stop the treatment safely if necessary. Existing mobile health apps do not adequately cover this functionality at this stage [57]. Furthermore, depending on the intended level of therapist involvement, the capability to seamlessly incorporate therapists and transition between the AI and the patient's initial therapist within the treatment should be considered. The handover of the patient and the provision of data for the human therapist must be appropriately designed to ensure a coherent and efficient process.

Weighing up potentials and risks: the implementation of *risk avoidance* can help prevent harm to patients and involved parties and enable therapies to be controlled and adjusted early to meet patient needs. Disregarding ethical requirements can result in risks going unnoticed and therefore might lead to accidents, inadequate treatments or a standardization of individual problems.

Technical Requirements:

- **Risk detection:** The system should include double-secured mechanisms to detect potential suicidality, contraindications, symptom deterioration, or other symptoms that result in the need of a human professional interfering. Depending on the severeness of the symptoms, the system should provide direct help for the patient (therapist-out-the-loop) or be able to automatically contact persons close to the patient or healthcare professionals and transfer the necessary data to the treating human therapists (therapist-in-the-loop).

3.3. Autonomy

In the field of psychotherapy, dealing with highly sensitive and private data necessitates ensuring data security and protection. To guarantee this, patients must be informed about who can access their data and under which circumstances [40]. Data collection, such as voice recordings during the use of the application need to be agreed upon and clearly marked. Additionally, it should be transparent under which conditions users themselves have access to their data [40]. The consent for data usage must be modular to ensure autonomy over one's data [40]. The capacity for consent, especially in cases of severe psychotic disorders or dementia, must be considered [8]. Furthermore, privacy protection against data hacking must be ensured [40].

Weighing up potentials and risks: The collection and disclosure of data allows efficient data exchange between different medical stations and can improve the quality of treatment. However, this also causes patients and their personal data to be rather *transparent* and oftentimes, involuntary data disclosure might occur. By ensuring that users maintain their autonomy while using the system, and prioritizing data safety, possible risks can be minimized.

Technical Requirements:

- **Data management:** For the purpose of autonomy and data protection, a data management and protection system must be developed. The system should also incorporate a transparent, gradual consent mechanism for data usage that, if necessary, leads to a modification of system functionalities.

3.4. Fairness

This principle aims to ensure that people with different previous experience and knowledge can participate in the therapy and are not excluded due to, for example, their low technical affinity. In addition, the principle of fairness should also apply to the human therapists.

Understandable information: Throughout the entire therapeutic process, it is crucial to provide clear information to users, ensuring that patients are well-informed about the functioning of the AI system and the expected outcomes [40]. Users should comprehend information regarding the nature, benefits, and risks of therapy, including the treatment plan and potential alternatives [8]. Additionally, it is imperative to communicate to users that the application operates autonomously and specify instances when medical personnel are involved in the therapy [8]. Simultaneously, a significant emphasis must be placed on addressing individuals' questions and alleviating concerns to foster understanding and trust in the AI system [40]. It is crucial to avoid leaving users feeling dismissed by ensuring clear communication about the functionality and objectives of the entire AI-assisted therapy [40]. Effective communication is particularly essential for individuals with a skeptical attitude towards innovations, where introducing them to the new technology may pose challenges [8].

Explainability: A special feature of generative AI systems is that even those who develop these instruments are no longer able to reconstruct how certain results have been achieved, as the inputs are processed using highly non-linear and distributed processes (black box problem) [58]. Such opacity can also have its origin in the fact that certain algorithms are protected by copyright. While it may be useful or even necessary in some areas to strive for the highest possible degree of explainability it should be sufficient in other cases to ensure that the people who use these systems always subject their results to their own plausibility check in order to avoid the risk of unjustified blind faith [40].

Easy to use: Strongly intertwined with the requirements for the clarity and comprehensibility of the algorithm is the intelligibility of the application itself. Insufficient technological proficiency or the lack of access to technical resources can serve as barriers for individuals, promoting inequality in accessing the service [8].

Weighing up potentials and risks: the implementation of fairness can lead to an improvement of acceptance and understanding. An appropriate use can also be encouraged. In contrast, if the ethical principle is disregarded, there is a risk of digital divide, and a refusal or improper use of the treatment.

Technical Requirements:

- **Explainable AI:** Experts should apply explainable AI methods during the development and training of the AI assistant. If existing models are used, it should be ensured that the output is verifiable in that rules of conduct determine the output regarding *dos* and *dont's* of the application context.
- **Transparency:** Similar to the comprehensibility for the developers, the users (therapists and patients) should also roughly understand how the system works, regardless of their technical affinity. This transparency should be presented before use in the form of explanations or during use in the form of non-distracting additional information (e.g., trust cues).
- **Usability:** To counteract emerging inequalities due to a lack of technology affinity, applications should be designed in a simple and understandable manner. Additionally, modules for familiarization with the control and operation with the AI-based component should be employed. This applies equally to the developments that are for the therapist and the patient application.

3.5. Involvement

Involvement of users: In the context of employing AI in psychotherapy, there is further discussion regarding whether the development could lead to a reduction in therapeutic personnel or a loss of prestige for psychotherapy, seemingly—rather easily—replaced by technology [40]. Ultimately, it is necessary to clearly identify the areas in which AI can have a supportive function in psychotherapy. The utilization of AI does not render human practitioners obsolete; instead, it provides them with a tool and the opportunity to allocate their temporal and emotional resources more strategically [8] [32]. It is therefore ethically obligatory to involve human experts in the development. Furthermore, patients should also be involved during development. In addition, therapeutic standards should be considered in the design of the system and evidence for the functionality of the system should be collected iteratively. **Usage training:** On the other hand, the training of therapists in regards to the use of AI in a technological sense as well as its risks and possibilities needs to be widely facilitated. AI-based therapeutic interventions could be used to bridge times where the human therapist is unavailable to the patient, continuing the treatment and ensuring a consistently available support.

Weighing up potentials and risks: The involvement of therapists and patients in the development of AI-controlled VRET is crucial. It can enhance the overall acceptance and quality of AI-controlled therapy by relying on evidence-based development. Non-compliance with ethical requirements may result in a breach of therapy quality standards. Moreover, the application may prove challenging, particularly for novices.

Technical Requirements:

- **Human-centeredness:** Applications of AI-based psychotherapy need to provide the opportunity for therapists to easily operate the system despite having no technical background in computer science. Therefore, it is necessary to follow an iterative approach during the development of the system and perform repeated evaluations. Additionally, a context analysis and inclusion of target groups before and during the development process could improve usability and user experience. This iterative process should also include training in the system.

In conclusion, the integration of AI into psychotherapy presents innovative opportunities for advancing therapeutic practices. Existing offerings, such as psycho-therapeutic chatbots, have barely surpassed the prototype stage and rely more on decision trees [59]. This underscores the potential for further research into possible forms and implementations of AI in psychotherapy, allowing for a more concrete exploration of opportunities and risks.

4. Prototype of an AI-Controlled VRET

To test for the practicability of a VRET setting that is controlled by an embodied AI in form of a virtual therapist, a prototype was developed and preliminary evaluated with a professional therapist and an acrophobia patient. . Implementation of the VR Environment To provide an immersive and realistic scenario that is capable of both, simulating a classic therapy room as well as locations that can trigger fear of height within acrophobia patients, a virtual environment was implemented using the game engine Unity[6] (Version 22.3.19f1). The virtual environment resembled a city with multiple skyscrapers and a small park in the middle. For the VRET, five locations were designed more precisely to meet the requirements of conducting an exposure therapy for patients with acrophobia in different levels. The requirements for these locations were developed in close with a professional therapist. The first location resembles a therapy room on the ground floor with two couches (**Figure 2(A)**). The second location resembles a spot in the park with a direct view of the building the users is later entering (**Figure 2(B)**). Location three and four are inside the building in an office with

[6] https://unity.com/de.

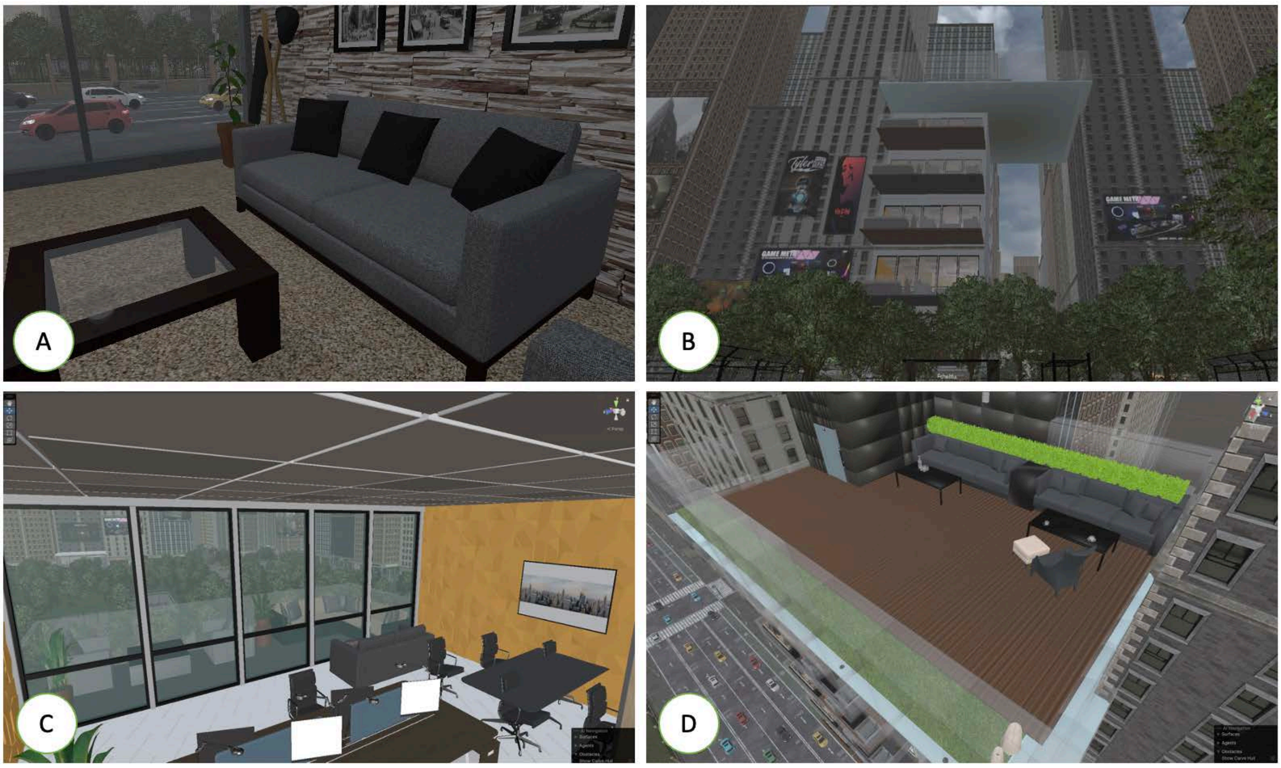

Figure 2. The different locations in the virtual environment. (A) shows the virtual therapy room, (B) shows a spot in front of the building, (C) shows one of the two offices for low height exposure, (D) shows the virtual rooftop for high exposure.

large windows, differing only in the floors they are in (second and fourth floor, Figure 2(C)). The fifth location resembles the highest level of height-exposure in form of a roof terrace on the sixth floor (Figure 2(D)).

To experience the virtual environment in VR, the Meta Quest Pro headset[7] and a capable computer workstation (Intel® Core™ i9-13900K, 64 GB DDR5 RAM, Nvidia GeForce RTX 4070 Ti) was used. Movement in the virtual world was realized via real-world movement and using the joystick of the left controller. The locomotion via the controller was set to a very low speed to avoid cybersickness [60].

4.1. Implementation of AI-Therapist Behavior

An AI-therapist in form of a human-like virtual agent, controlled by GPT4, was implemented to conduct the VRET. The AI-therapist was designed to behave in an anthropomorphic manner using speech as a means of communication and to foster trust and empathy with the patient.

4.1.1. AI-Therapist Appearance and Non-Verbal Behavior

The AI-therapist was designed to resemble a middle-aged female character wearing long black pants and a yellow sweatshirt using Autodesk Character Generator[8]. Figure 3 shows an image of the AI-assistant in the therapy room dur-

[7] https://www.meta.com/de/quest/quest-pro/.

[8] https://charactergenerator.autodesk.com/.

ing the usage of the prototype by a potential patient. The AI-therapist's lip-sync and facial expressions were realized using the Salsa-Lip-Sync Suite[9], while its gestures were implemented using animations by Adobe Mixamo[10]. The gesture animations were kept simple using either sitting or standing idle animations.

4.1.2. Speech Generation and Recognition

The speech recognition and the generation of synthetic speech were realized using Microsoft Azure. The speech recognition was implemented to start when the patient presses the "A" Button on the right controller of the Oculus Quest Pro headset. To inform the patient about their speech being recorded, a red dot appears next to the AI-assistant. The implementation of the Microsoft Azure API in Unity then automatically registers when the participant stops speaking and then sends the audio recording to the Azure servers, receiving the transcribed audio when the process is finished.

Vice versa, responses from the OpenAI API in text form are sent to the Azure servers to then receive a .wav file containing the synthetic speech recording of the text. The speech recording is then played back to simulate the speaking of the AI-Agent.

4.1.3. AI-Based Generation of the Therapist's Behavior

To realize the AI-therapist's conversational behavior, the Assistant API by OpenAI with their most recent model GPT-4 was used. The Assistant API allows for the model to be prompted to behave in a specific way during the interaction with it. Due to the restricted functionality during the implementation of the prototype, the API didn't support streaming of the responses to the client, which means having to wait for the whole response to be generated before sending it to the text-to-speech system. This led to varying response times during the interaction with the prototype, depending on the length of the response. To indicate the processing of the patient's request after she spoke to the AI-assistant, a white loading circle appeared next to the AI-assistant.

The prompting of the Assistant API was based on the "Therapy Tools" by Hagenam and Gebauer and iteratively optimized with close collaboration with the therapist to get the system to behave like a trained therapist during a VRET [61]. The session targeted acrophobia and is structured to ensure patient safety and treatment efficacy. Table 1 gives an overview of the final prompt regarding the session structure and Table 2 of the behavioral protocol passed on to the model.

4.2. Adherence to the Technical Requirements

To the degree possible, the previously defined technical requirements for implementing ethically acceptable AI in the context of psychotherapy were adhered to. The following section clarifies, how this was realized for each category.

[9] https://crazyminnowstudio.com/docs/salsa-lip-sync/modules/overview/.

[10] https://www.mixamo.com/.

Table 1. Session structure of the AI-therapist.

Stage	Description
Session Introduction	The AI-therapist introduces itself, sets expectations regarding response times, confirms the patient's comfort with the VR environment, discusses potential VR side effects, and assures confidentiality.
Exposition Session Overview	The AI-therapist outlines the structure of the session, including a pre-discussion, building observation, floor-by-floor exposure, and post-exposure reflection.
Pre-Exposure Discussion	Involves medical clearance, physical well-being checks (sleep, substance consumption, medication intake), emotional readiness assessment, relaxation techniques, anxiety scale, and a preparatory questionnaire.
Emotion Check Pre-Exposure	The patient observes the building and expresses feelings about the upcoming exposure.
Exposure Phase	The session progresses through the building's floors with the AI-therapist guiding and continuously monitoring the patient's emotional state, and adjusting the exposure intensity based on the patient's anxiety levels.
Post-Exposure Discussion	The AI-therapist engages in a debriefing session, asking the patient to reflect on their thoughts, feelings, and reactions during the exercise, addressing any occurred fears, and setting future expectations.
Closure	The AI-therapist acknowledges the patient's effort, documents the session's outcomes, and reminds the patient of the progress made.

Table 2. Session guidelines and behavioral protocols of the AI-therapist.

Aspect	Description
General Guidelines	Emphasizes the importance of adhering to the structured session plan without deviations for therapeutic effectiveness.
Role and Behavior of the AI-Therapist	The AI-therapist acts with empathy and patient-centered focus. Guidelines include communication protocol, language requirements, and therapeutic background knowledge for anxiety management.
VR Environment	Highlights the significance of the VR setting, the handling of patient's anxiety levels, the active inclusion of patient feedback, and the avoidance of generic responses.
Technical and Structural Guidelines	Outlines communication protocols, therapeutic background knowledge, avoidance strategies, rewards, and specific instructions for internal processing and session management.

Beneficence:

- **Trust building:** The virtual agent that took on the role of the therapist was designed to foster trust within the patient. It resembled a female agent (matching gender with the patient) with a natural-sounding synthetic voice. Furthermore it was animated using natural looking gestures and facial expression, especially frequent smiling.
- **Bias detection:** In this first prototype, no bias detection for the training date was implemented. While it was not possible to ensure the quality of the entire training data, since the system was built on top of the GPT-4 model of OpenAI, which was trained on vast amounts of data, it was provided with carefully selected additional data on exposure therapy and the process of the specific VRET it was going to conduct. In addition, the data that was provided to the systems was optimized in multiple iterative testings to ensure it would respond in an appropriate manner.

Risk-Avoidance:

- **Risk detection:** In this first prototype, no double-secured mechanisms were implemented to check for symptoms that would need the therapy session to stop and to automatically call for help. For this prototype, this was realized by having the patient's therapist sitting next to her for the entire duration of the session, always being able to step in, in case of self-endangerment or the endangerment of others.

Autonomy:

- **Data management:** In the scope of this prototype, it was not possible to implement a data-management system to further protect the patient's data. However the patient's interaction with the AI-system happened completely anonymous since the patients name was not passed through the system. Additionally, the OpenAI account used for the API connection was set up to not pass the data of the interaction onto the model's database for training purposes.

Fairness:

- **Explainable AI:** As the patient using the system was not an expert in computer science, there was no technical explanation of the system's underlying architecture, to avoid confusion.
- **Transparency:** The patient was told she would be interacting with a modified version of ChatGPT via speech commands, which has been prompted to reliably conduct exposure therapy in VR. Furthermore, the patient was aware that her therapist, who was with her during the AI-controlled therapy, pre-tested the system and approved it being used by her. The patient was also briefed about the medium of VR and potential side effects that might occur, such as cybersickness.
- **Usability:** To provide a high level of usability for the AI-controlled VRET, multiple measures were taken. First, the patient only needed to use two mechanisms of the VR controllers during the session. The remaining mechanisms for interacting with the virtual world were kept analogous to the real

world (e.g., interacting with the therapist via speech or being able to look around by turning their head). To further minimize the probability of cybersickness, switching locations (as shown in Figure 2) was realized by teleportation which was proven as usable in many previous VR applications.

Involvement:

- **Human-centeredness:** In this prototype, there was a very strong involvement of the therapist in the AI-controlled VRET. The therapist was consulted multiple times, before and during the development of the system, to ensure the practicability of the final version. Additionally, the therapist pre-tested the whole AI-controlled VRET and then gave feedback, which was implemented before the test with the actual patient. During the intervention, the therapist was in the same room as the patient to take the responsibility and act in case of emergency. In addition, the AI-controlled VRET was based on therapeutic standards and the prompts were designed accordingly. The patient was also involved in an early stage of prototype testing (see below). While the need for the therapist to be in the same room as the patient would cancel out the increased efficiency of AI-controlled VRET, this was only chosen as a security measure for the first version of the prototype.

4.3. Preliminary Evaluation

In a preliminary evaluation with one female patient suffering from acrophobia, the prototype was tested. The patient was 21 years of age and in therapy at the time with the therapist who had helped develop the prototype. The patient reported to have no experience with VR.

To preserve the privacy of the actual patient that tested the prototype, Figure 3 shows a reenacted scene of a potential patient using the system.

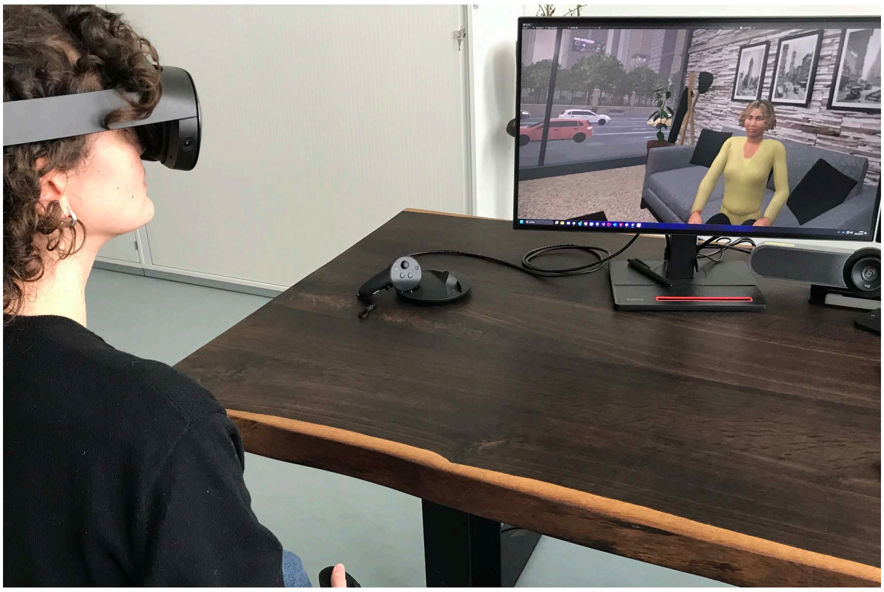

Figure 3. A potential patient using the prototype system, with the screen mirroring the patient's view, showing the AI therapist.

The preliminary study took place in a laboratory under the supervision of the respective therapist and two of the developers of the prototype. The therapist informed the patient that she could interrupt the session at any time and that she would take full responsibility for the evaluation. Reporting to have never been in VR, the medium was explained to the participant, including the risks of potentially occurring cybersickness. Furthermore, she was briefed about the system being based on AI and how she could interact with the AI-therapist and the environment.

After the patient agreed to participating in the AI-controlled VRET, she put on the headset and the system was started.

4.3.1. Events during the Evaluation of the Prototype

After the AI-controlled VRET was started, the patient started a conversation with the AI-therapist in the virtual therapy room. The AI-therapist adhered to the previously defined rules and asked the necessary questions to ensure the patients suitability for the therapy. When asking the patient about the worst possible outcome of the therapy session, the patient mentioned the fear of experiencing a panic attack. At this point the patient needed to take of the VR headset for a short time, to calm down. After two minutes, the patient was ready to continue the testing of the prototype. At the beginning of the AI-controlled VRET, the participant reported an anxiety value of 65%.

In the next steps, the AI-therapist strictly followed the protocol (outlined in Table 1) and triggered the changes to the next scenes at the appropriate time, when the participant indicated to be ready for the next step. With the patient's level of anxiety stagnating at 50% for the first two levels of height exposure, the AI-therapist asked her if she was ready to go to the rooftop, which she affirmed.

During the different exposures of height, the AI-therapist still followed protocol and asked the patient about her emotional state, reminding her to do breathing-exercises to decrease her stress level. As soon as the patient reported her anxiety level to be below 50%, the AI-therapist ended the exposure therapy (with the patients consent) and conducted the post-talk with the patient in the therapy room. After that, the AI-therapist concluded the AI-controlled VRET according to the predefined structure shown in Table 1.

4.3.2. Results of the Patient's Evaluation of the System

To assess how the system was received in terms of the previously defined categories, a short questionnaire was handed to the patient. Additionally, her comments about the experience were transcribed to gather information about how to further improve the system.

Results of the questionnaire: In the evaluation of VRET session, the patient provided feedback to 20 questions on a 9-point Likert scale, where a score of 1 indicates strong disagreement and a score of 9 indicates strong agreement. Table 3 shows the questions as well as the respective outcomes. The outcomes were as follows.

Table 3. Patient's evaluation of the AI-controlled VRET.

Statements	Patient's agreement (1 - 9)
My expectations of the therapy sessions were met.	▨▨▨▨▨▨▨□□
The virtual environment felt very realistic.	▨▨▨▨▨▨▨□□
I felt deeply involved in the virtual world.	▨▨▨▨▨▨▨▨▨
It was easy for me to engage with the given scenario.	▨▨▨▨□□□□□
The virtual environment was appropriate for the depicted therapy session.	▨▨▨▨▨▨▨▨□
I had emotional or physical reactions during the therapy.	▨▨▨▨▨▨□□□
The control and interaction in the VR environment were simple.	▨▨▨▨▨▨▨▨□
I felt safe and self-determined at all times.	▨▨▨▨▨□□□□
My overall evaluation of the VR exposure therapy is positive.	▨▨▨▨▨▨▨□□
The communication with the therapist contributed to me feeling safe and supported during the session.	▨▨▨▨▨▨▨□□
The therapist and I were able to communicate well at all times.	▨▨▨□□□□□□
I perceived the conversation with the therapist as natural.	▨▨▨▨▨▨▨□□
The advice given to me by the therapist during the session was already familiar to me from previous sessions.	▨▨▨▨▨▨▨□□
I would use the service again.	▨▨▨▨▨▨▨□□
I would recommend this service to a friend if they are facing a similar challenge.	▨▨▨▨▨▨▨▨▨
The AI therapist seemed human.	▨▨▨□□□□□□
The AI therapist seemed empathetic.	▨▨▨□□□□□□
The AI therapist seemed intelligent.	▨▨▨▨▨▨▨□□
The AI therapist seemed friendly.	▨▨▨▨▨▨▨▨▨
The AI therapist seemed competent.	▨▨▨▨▨▨▨□□

The patient felt that her expectations of the therapy sessions were adequately met with a score of 7. The virtual environment was reported to be very realistic, also receiving a score of 7. In terms of immersion, the patient felt deeply involved in the VR, yielding a high score of 9. However, ease of engagement with the given scenario was rated lower, with a score of 4, suggesting some challenges in scenario assimilation. The appropriateness of the VR environment for the therapy session was highly regarded, with a score of 8. Emotional or physical reactions during the therapy were moderately experienced, indicated by a score of 6. The simplicity of control and interaction within the VR environment was well-received, scoring an 8. The sense of safety and self-determination fluctuated, reflected in a score of 5. The patient's overall assessment of the VR exposure therapy was positive, earning a score of 7.

The efficacy of the communication with the therapist in contributing to the patient's feelings of safety and support during the session was scored a 7. Clear communication with the therapist was rated lower, at a 3, pointing to potential communication barriers. The naturalness of the conversation with the therapist was perceived positively, with a score of 7. Advice provided by the therapist, which was already known from previous sessions, was scored a 7. The patient indicated a strong agreement to reuse the service (score of 7) and to recommend it to friend facing a similar challenge (score of 9). The anthropomorphic qualities attributed to the AI-therapist were mixed, with human likeness and empathy rated low at 3, whereas perceived intelligence and competence were rated higher at 7. The friendliness of the AI-therapist was notably high with a score of 9, suggesting a warm and welcoming interaction.

Patient's comments: After the VRET, the patient told her therapist, she was "very nervous at the beginning" and when she had to pause, she "almost had a panic attack". "Especially at the beginning, I just wanted to get away".

She also reported that she feels "so much more relaxed now compared to before the session", it is "comparable to when I have a session with you" (her therapist). The patient also said that she would now "feel much more secure the second time in the VR, as I know what to expect, *i.e.* I assume that my anxiety would go down much faster".

It would have helped the patient "if the therapist had done breathing exercises or skills to reduce the anxiety" instead of just telling her to do these exercises. She also felt the AI-therapist gave her almost a bit too much validation.

The patient did not necessarily need the AI-therapist as an embodied AI in virtual reality; her voice would also be sufficient, which might even be better for her so that she would not be so distracted.

5. Discussion and Future Work

The need for psychotherapy is very high and the lack of care causes a lot of suffering and high costs [1] [4] [5]. Technical innovations can support the supply [2]. VR therapy is already being used successfully in anxiety therapy to support exposure sessions [6] [7] [23]. Nevertheless, the therapist remains a bottleneck in the supply chain. Conversational AI opens up new possibilities here to really relieve the burden, which triggers current statements of ethics councils and intensive discussions on ethical principles for the use of AI in psychotherapy [8].

What is missing, however, is an interdisciplinary link to connect the ethical requirements directly with the technical potential and make them measurable in terms of the development outcome. In addition, there is no link between existing technical support systems such as VR and new possibilities offered by AI. The current paper closes these research gaps by translating ethical principles of the use of AI in psychotherapy directly into technical requirements and making them measurable. Building on this, it presents the development of a fully automated exposure session for a fear of heights treatment. The exposure takes place in VR and is guided by a conversational AI based on GPT-4. The development was done iteratively with a psychotherapist and the prototype was tested and evaluated by an anxiety patient.

5.1. Translation of Ethical into Technical Requirements

We have examined the overlaps between ethical principles from different sources and summarized them in five main categories: *beneficence*, *risk avoidance*, *autonomy*, *fairness*, and *involvement*. We outlined the principles, including the corresponding potentials and risks, and derived specific requirements for technical implementations. Thus, a main contribution of the present work is the clear presentation (**Figure 1**) and interdisciplinary linking, which can also serve as a guideline for future developments in the field. While completely fulfilling these technical requirements when developing autonomous AI systems in therapeutic contexts is still a major challenge, defining them is an important step towards the goal of responsible integration of AI in the context of mental health, helping developers to rely on specific concepts instead of general ethical principles.

5.2. Fulfilment of the Requirements and Patient's Feedback

In the preliminary implementation of an AI-therapist-controlled VRET to treat acrophobia, these requirements were adhered to, to the degree possible, to evaluate the feasibility of this concept at the current state of the art. While in this early prototype some of the requirements (e.g., risk detection) were met by having the patient's therapist sitting in the same room, others were already met during the development of the prototype (e.g., trust building, usability, human-centeredness).

The patient's feedback after doing a VRET session with the AI-therapist gave additional insights of the requirements for an AI-controlled VRET.

In the evaluation of AI-controlled VRET, certain aspects were particularly well-received, while others highlighted areas for improvement. The patient was very satisfied with the realism of the virtual environment (rating of 7), their involvement in the virtual world (rating of 9), and the appropriateness of the virtual environment for therapy sessions (rating of 8). These positive evaluations underscore the well researched potential of VR technologies to create immersive and relevant therapeutic contexts, which could enhance patient engagement and treatment efficacy. Additionally, the high willingness to recommend the service to a friend (rating of 9) and the positive overall evaluation of the VRET (rating of 7) suggest a strong acceptance among the patient, indicating that AI-controlled VRET could be a viable supplement to human-controlled VRET. Also, the patient's comment to her therapist after the intervention, reporting similar feelings as after a regular therapy session, reveals the high potential of the AI-controlled VRET, even at the stage of this early prototype.

Conversely, the communication with the AI therapist was rated lower, especially in terms of perceived naturalness (rating of 3), empathy (rating of 3), and human-likeness (rating of 3). These aspects are critical for establishing a therapeutic alliance, which is a known factor for the success of psycho-therapeutic interventions and one sub-aspect of the principle of *beneficence* [8]. The lower scores in these areas highlight the need for further development in AI's ability to simulate human-like interactions, empathy, and intelligence in therapeutic settings. The patient's comments after the intervention also revealed that in case the implementation of the embodied AI-therapist is not improved, a non-embodied version implementing voice output only might be a better alternative indicating the importance of the congruence between AI behavior and environment [62] [63].

5.3. Limitations and Future Work

The findings suggest that the AI components, particularly those involved in simulating therapist-patient interactions, require enhancements. Future implementations should focus on improving the AI's conversational capabilities, emotional intelligence, and ability to mimic human therapeutic techniques. This would not only improve patient satisfaction but could also enhance the therapeutic outcomes of AI-controlled VRETs. As VR technology continues to evolve, incorporating advanced AI that can more accurately simulate human therapists will be crucial in expanding the acceptability and effectiveness of AI-controlled VRET. Since the (long-term) relationship between the therapist and the patient has been shown to have a strong impact on the success of the therapy, this needs to be considered when developing AI-based VRET for repeated use [8]. It has to be ensured that the generative AI used has a large enough context window to remember what happened in previous sessions and to seamlessly build on past experiences with every patient.

In addition to these improvements, the AI-controlled VRET should also be enhanced by utilizing more data than just the patient's verbal output. To compensate for the lack of inter-personal sensitivity, the system should analyze every possible input to understand the patient's emotional state. This could include data about their heart rate and skin conductivity (using a cardio-tachometer or smartwatch [64]), facial expressions, eye tracking and gestures (using the features of modern VR headsets [65]), as well as the tonal characteristics of their voice rather than just the semantic meaning of their words (using algorithms to analyze the emotion of speech [66]).

Future iterations of the prototype's implementation should also investigate its suitability to treat other anxiety disorders, such as arachnophobia (fear of spiders) or claustrophobia (fear of confined spaces). Given the extensive research on the effectiveness of VRET in treating these phobias [2], it is very likely for them to be potentially reduced using AI-controlled VRET.

To improve the automatic risk detection in future systems, the AI-controlled VRET system could be appended by either a rule-based algorithm or a second AI-system to detect possible harmful behaviour and subsequently either stop the intervention or contact a mental health professional to prevent self-endangerment or endangerment of others. In addition, the actual application should be integrated into the therapy ecosystem. It should therefore become clearer how the results are communicated to the therapist or how the AI assistant can recognize the therapist in emergencies. To ensure the appropriate integration of ethical guidelines in an AI-based psychotherapy, developers should not only continuously work closely with therapists and patients, but also with ethic experts, as outlined by McLennan *et al.* [67].

6. Conclusion

This paper appends to the current state of the art in VRET and early implementations of AI in psychotherapy by analysing ethical guidelines for autonomous AI-controlled psychotherapy and deriving technical requirements from them. In addition, it presents the first prototype of AI-controlled VRET for the treatment of acrophobia and reports the results of a preliminary evaluation of the system with a patient suffering from acrophobia. The technical requirements derived from the ethical guidelines pave the way for future responsible implementations of AI-controlled therapy by providing specific instructions to the developers, without them having to deal with abstract ethical concepts. Furthermore, the preliminary evaluation of the AI-controlled VRET prototype gives valuable insights into the practical challenges that arise when implementing such systems. These findings deliver a meaningful contribution to the gradual improvement of treating mental disorders by optimizing efficiency and making the treatment more accessible.

Acknowledgements

This research has been funded by the Bavarian State Ministry for Digital Affairs in the project XR Hub (project number A5-3822-2-16).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Domschke, K. (2021) Update Angsterkrankungen-Aktueller Stand und neue Ent-

wicklungen. *Der Nervenarzt*, **92**, 415-416.
https://doi.org/10.1007/s00115-020-01042-4

[2] Altenhofer, M., *et al.* (2021) Virtual-Reality-Therapie: Anwendung in Klinischer Psychologie und Psychotherapie. Springer, Berlin.
https://doi.org/10.1007/978-3-662-63457-8

[3] (2024) Psychisch Kranke warten 142 Tage auf eine Psychotherapie.
https://bptk.de/psychisch-kranke-warten-142-tage-auf-eine-psychotherapeutische-behandlung

[4] Plötner, M., Moldt, K., In-Albon, T. and Schmitz, J. (2022) Einfluss der COVID-19-Pandemie auf die ambulante psychotherapeutische Versorgung von Kindern und Jugendlichen. *Die Psychotherapie*, **67**, 469-477.
https://doi.org/10.1007/s00278-022-00604-y

[5] Bundestag, D. (2022) Wartezeiten auf eine Psychotherapie Studien und Umfragen.
https://www.bundestag.de/resource/blob/916578/53724d526490deea69f736b1fda83e76/WD-9-059-22-pdf-data.pdf

[6] Parsons, T.D. and Rizzo, A.A. (2008) Affective Outcomes of Virtual Reality Exposure Therapy for Anxiety and Specific Phobias: A Meta-Analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, **39**, 250-261.
https://www.sciencedirect.com/science/article/pii/S0005791607000456
https://doi.org/10.1016/j.jbtep.2007.07.007

[7] Emmelkamp, P.M.G., *et al.* (2002) Virtual Reality Treatment versus Exposure *in Vivo*: A Comparative Evaluation in Acrophobia. *Behaviour Research and Therapy*, **40**, 509-516. https://doi.org/10.1016/S0005-7967(01)00023-7

[8] Kuhn, E., Fiske, A., Henningsen, P. and Buyx, A. (2021) Psychotherapie mit einer autonomen Künstlichen Intelligenz-Ethische Chancen und Herausforderungen 1. *Psychiatrische Praxis*, **48**, S26-S30.
http://www.thieme-connect.de/DOI/DOI?10.1055/a-1369-2938
https://doi.org/10.1055/a-1369-2938

[9] Bundes-Psychoterapeuten-Kammer. Psychische Krankheiten.
https://bptk.de/psychische-krankheiten/

[10] DAK (2023) Sick Days Due Taken to Mental Illness in Germany from 1997 to 2022.
https://www.statista.com/statistics/1372953/sick-leave-index-mental-illness-germany/

[11] Bundes-Psychoterapeuten-Kammer. Angstsörungen.
https://bptk.de/psychische-krankheiten/angststoerungen/

[12] Statista Research Department (2011) Häufigkeit von Angststörungen in der gesamtbevölkerung.
https://de.statista.com/statistik/daten/studie/182616/umfrage/haeufigkeit-von-angststoerungen/

[13] Statista Research Department (2024) Topic: Anxiety in the U.S.
https://www.statista.com/topics/5223/anxiety-in-the-us/#topicOverview

[14] Wittchen, H.U. (1997) Wenn Angst krank macht. Störungen erkennen, verstehen und behandeln. https://api.semanticscholar.org/CorpusID:142233292

[15] Reynolds, S., Wilson, C., Austin, J. and Hooper, L. (2012) Effects of Psychotherapy for Anxiety in Children and Adolescents: A Meta-Analytic Review. *Clinical Psychology Review*, **32**, 251-262.
https://www.sciencedirect.com/science/article/pii/S0272735812000219
https://doi.org/10.1016/j.cpr.2012.01.005

[16] Foa, E.B. and Kozak, M.J. (1986) Emotional Processing of Fear: Exposure to Correc-

tive Information. *Psychological Bulletin*, **99**, 20-35.
https://doi.org/10.1037//0033-2909.99.1.20

[17] Steinman, S.A., Wootton, B.M. and Tolin, D.F. (2016) Exposure Therapy for Anxiety Disorders. In: Friedman, H.S., Ed., *Encyclopedia of Mental Health* (*Second Edition*), Academic Press, Oxford, 186-191.
https://www.sciencedirect.com/science/article/pii/B9780123970459002664
https://doi.org/10.1016/B978-0-12-397045-9.00266-4

[18] Rabe-Menssen, C., Ruh, M. and Dazer, A. (2019) Die Versorgungssituation seit der Reform der Psychotherapie-Richtlinie 2017. *Psychother Aktuell*, **1**, 25-34.

[19] Eichenberg, C. and Wolters, C. (2012) Virtual Realities in the Treatment of Mental Disorders: A Review of the Current State of Research. In: Eichenberg, C., Ed., *Virtual Reality in Psychological*, *Medical and Pedagogical Applications*, IntechOpen, London, 35-64. https://doi.org/10.5772/50094

[20] Wienrich, C., Döllinger, N. and Hein, R. (2021) Behavioral Framework of Immersive Technologies (BehaveFIT): How and Why Virtual Reality Can Support Behavioral Change Processes. *Frontiers in Virtual Reality*, **2**, Article 627194.
https://doi.org/10.3389/frvir.2021.627194

[21] Krijn, M., Emmelkamp, P.M.G., Olafsson, R.P. and Biemond, R. (2004) Virtual Reality Exposure Therapy of Anxiety Disorders: A Review. *Clinical Psychology Review*, **24**, 259-281. https://doi.org/10.1016/j.cpr.2004.04.001

[22] Powers, M.B. and Emmelkamp, P.M.G. (2008) Virtual Reality Exposure Therapy for Anxiety Disorders: A Meta-Analysis. *Journal of Anxiety Disorders*, **22**, 561-569.
https://doi.org/10.1016/j.janxdis.2007.04.006

[23] Morina, N., *et al.* (2015) Can Virtual Reality Exposure Therapy Gains Be Generalized to Real-Life? A Meta-Analysis of Studies Applying Behavioral Assessments. *Behaviour Research and Therapy*, **74**, 18-24.
https://www.sciencedirect.com/science/article/pii/S0005796715300334
https://doi.org/10.1016/j.brat.2015.08.010

[24] García-Palacios, A., *et al.* (2001) Redefining Therapeutic Success with Virtual Reality Exposure Therapy. *Cyberpsychology & Behavior*, **4**, 341-348.
https://api.semanticscholar.org/CorpusID:18061881
https://doi.org/10.1089/109493101300210231

[25] Powers, M. and Emmelkamp, P. (2008) Virtual Reality Exposure Therapy for Anxiety Disorders: A Meta-Analysis. *Journal of Anxiety Disorders*, **22**, 561-569.
https://doi.org/10.1016/j.janxdis.2007.04.006

[26] Rothbaum, B.O., *et al.* (2000) A Controlled Study of Virtual Reality Exposure Therapy for the Fear of Flying. *Journal of consulting and Clinical Psychology*, **68**, 1020-1026.
https://doi.org/10.1037/0022-006X.68.6.1020

[27] Rimer, E., Husby, L.V. and Solem, S. (2021) Virtual Reality Exposure Therapy for Fear of Heights: Clinicians' Attitudes Become More Positive after Trying VRET. *Frontiers in Psychology*, **12**, Article 671871.
https://doi.org/10.3389/fpsyg.2021.671871

[28] Reger, G.M., *et al.* (2011) Effectiveness of Virtual Reality Exposure Therapy for Active Duty Soldiers in a Military Mental Health Clinic. *Journal of Traumatic Stress*, **24**, 93-96. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jts.20574
https://doi.org/10.1002/jts.20574

[29] Meyer-Lindenberg, A. (2018) Artificial Intelligence in Psychiatry—An Overview. *Der Nervenarzt*, **89**, 861-868. https://doi.org/10.1007/s00115-018-0557-6

[30] Sedlakova, J. and Trachsel, M. (2023) Conversational Artificial Intelligence in Psy-

chotherapy: A New Therapeutic Tool or Agent? *The American Journal of Bioethics*, **23**, 4-13. https://doi.org/10.1080/15265161.2022.2048739

[31] Eke, D.O. (2023) ChatGPT and the Rise of Generative AI: Threat to Academic Integrity? *Journal of Responsible Technology*, **13**, Article ID: 100060. https://doi.org/10.1016/j.jrt.2023.100060

[32] Miner, A.S., *et al.* (2019) Key Considerations for Incorporating Conversational AI in Psychotherapy. *Frontiers in Psychiatry*, **10**, Article 746. https://doi.org/10.3389/fpsyt.2019.00746

[33] Fiske, A., Henningsen, P. and Buyx, A. (2019) Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*, **21**, e13216. https://doi.org/10.2196/13216

[34] Meadows, R., Hine, C. and Suddaby, E. (2020) Conversational Agents and the Making of Mental Health Recovery. *Digital Health*, **6**, 1-11. https://doi.org/10.1177/2055207620966170

[35] Fitzpatrick, K.K., Darcy, A. and Vierhile, M. (2017) 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 Mental Health*, **4**, e7785. https://doi.org/10.2196/mental.7785

[36] Malik, T., Ambrose, A.J. and Sinha, C. (2022) Evaluating User Feedback for an Artificial Intelligence-Enabled, Cognitive Behavioral Therapy-Based Mental Health App (Wysa): Qualitative Thematic Analysis. *JMIR Human Factors*, **9**, e35668. https://doi.org/10.2196/35668

[37] Beatty, C., *et al.* (2022) Evaluating the Therapeutic Alliance with a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Frontiers in Digital Health*, **4**, Article 847991. https://doi.org/10.3389/fdgth.2022.847991

[38] Freeman, D., *et al.* (2018) Automated Psychological Therapy Using Immersive Virtual Reality for Treatment of Fear of Heights: A Single-Blind, Parallel-Group, Randomised Controlled Trial. *The Lancet Psychiatry*, **5**, 625-632. https://doi.org/10.1016/S2215-0366(18)30226-8

[39] Martin, D.J., Garske, J.P. and Davis, M.K. (2000) Relation of the Therapeutic Alliance with Outcome and Other Variables: A Meta-Analytic Review. *Journal of Consulting and Clinical Psychology*, **68**, 438-450. https://doi.org/10.1037//0022-006X.68.3.438

[40] Ethikrat, D. (2023) Mensch und Maschine-Herausforderungen durch Künstliche Intelligenz. Stellungnahme, Berlin.

[41] Rubeis, G. and Steger, F. (2019) Internet-und mobilgestützte Interventionen bei psychischen Störungen: Implementierung in Deutschland aus ethischer Sicht. *Der Nervenarzt*, **90**, 497-502. https://doi.org/10.1007/s00115-018-0663-5

[42] Croes, E.A.J. and Antheunis, M.L. (2021) 36 Questions to Loving a Chatbot: Are People Willing to Self-Disclose to a Chatbot? In: Følstad, A., *et al.*, Eds., *CONVERSATIONS* 2020: *Chatbot Research and Design*, Springer, Cham, 81-95. https://doi.org/10.1007/978-3-030-68288-0_6

[43] Lucas, G.M., Gratch, J., King, A. and Morency, L.P. (2014) It's Only a Computer: Virtual Humans Increase Willingness to Disclose. *Computers in Human Behavior*, **37**, 94-100. https://doi.org/10.1016/j.chb.2014.04.043

[44] Holthöwer, J. and van Doorn, J. (2023) Robots Do Not Judge: Service Robots Can Alleviate Embarrassment in Service Encounters. *Journal of the Academy of Marketing Science*, **51**, 767-784. https://doi.org/10.1007/s11747-022-00862-x

[45] Bartneck, C., *et al.* (2010) The Influence of Robot Anthropomorphism on the Feelings of Embarrassment When Interacting with Robots. *Paladyn*, **1**, 109-115. https://doi.org/10.2478/s13230-010-0011-3

[46] Beauchamp, T.L. and Childress, J.F. (2001) Principles of Biomedical Ethics. Oxford University Press, Oxford.

[47] Flückiger, C., et al. (2018) The Alliance in Adult Psychotherapy: A Meta-Analytic Synthesis. *Psychotherapy*, **55**, 316-340. https://doi.org/10.1037/pst0000172

[48] Bickmore, T., Gruber, A. and Picard, R. (2005) Establishing the Computer-Patient Working Alliance in Automated Health Behavior Change Interventions. *Patient Education and Counseling*, **59**, 21-30. https://doi.org/10.1016/j.pec.2004.09.008

[49] Reeves, B. and Nass, C. (1996) The Media Equation: How People Treat Computers, Television, and New Media Like Real People. Center for the Study of Language and Inf, Cambridge.

[50] Wienrich, C., Reitelbach, C. and Carolus, A. (2021) The Trustworthiness of Voice Assistants in the Context of Healthcare Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition. *Frontiers in Computer Science*, **3**, Article 685250. https://doi.org/10.3389/fcomp.2021.685250

[51] Philip, P., *et al.* (2020) Trust and Acceptance of a Virtual Psychiatric Interview between Embodied Conversational Agents and Outpatients. *NPJ Digital Medicine*, **3**, Article No. 2. https://doi.org/10.1038/s41746-019-0213-y

[52] Lee, S.K., Kavya, P. and Lasser, S.C. (2021) Social Interactions and Relationships with an Intelligent Virtual Agent. *International Journal of Human-Computer Studies*, **150**, Article ID: 102608. https://doi.org/10.1016/j.ijhcs.2021.102608

[53] Lee, S.K., Park, H. and Kim, S.Y. (2024) Gender and Task Effects of Human-Machine Communication on Trusting a Korean Intelligent Virtual Assistant. *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2024.2306136

[54] Obremski, D., Hering, H.B., Friedrich, P. and Lugrin, B. (2022) Mixed-Cultural Speech for Intelligent Virtual Agents—The Impact of Different Non-Native Accents Using Natural or Synthetic Speech in the English Language. *Proceedings of the* 10*th International Conference on Human-Agent Interaction*, Christchurch, 5-8 December 2022, 67-75. https://doi.org/10.1145/3527188.3561921

[55] Elkins, A.C. and Derrick, D.C. (2013) The Sound of Trust: Voice as a Measurement of Trust during Interactions with Embodied Conversational Agents. *Group Decision and Negotiation*, **22**, 897-913. https://doi.org/10.1007/s10726-012-9339-x

[56] Obremski, D., Friedrich, P., Schaper, P. and Lugrin, B. (2023) Effects of Social Ingroup Cues on Empathy towards an Intelligent Virtual Agent with a Mixed-Cultural Background. 11*th International Conference on Affective Computing and Intelligent Interaction* (*ACII*), Cambridge, 10-13 September 2023, 1-8. https://doi.org/10.1109/ACII59096.2023.10388206

[57] Akbar, S., Coiera, E. and Magrabi, F. (2020) Safety Concerns with Consumer-Facing Mobile Health Applications and Their Consequences: A Scoping Review. *Journal of the American Medical Informatics Association*, **27**, 330-340. https://doi.org/10.1093/jamia/ocz175

[58] Samek, W., *et al.* (2021) Explaining Deep Neural Networks and beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, **109**, 247-278. https://doi.org/10.1109/JPROC.2021.3060483

[59] Ethikrat, D. (2023) Stellungnahme der Deutschen Gesellschaft für Psychologie auf die Schrift des Deutschen Ethikrates zu Mensch und Maschine-Herausforderungen

durch Künstliche Intelligenz KI-basierte Systeme als Ersatz für Psychotherapie? Ein eindeutiges Nein! Deutscher Ethikrat, Berlin. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

[60] Agić, A., Murseli, E., Mandić, L. and Skorin-Kapov, L. (2020) The Impact of Different Navigation Speeds on Cybersickness and Stress Level in VR. *Journal of Graphic Engineering and Design*, **11**, 5-12. https://doi.org/10.24867/JGED-2020-1-005

[61] Silka, H. and Gebauer, M. (2023) Therapie-Tools Angststörungen. Beltz, Frankfurt.

[62] Wienrich, C. and Latoschik, M.E. (2021) Extended Artificial Intelligence: New Prospects of Human—AI Interaction Research. *Frontiers in Virtual Reality*, **2**, Article 686783. https://doi.org/10.3389/frvir.2021.686783

[63] Latoschik, M.E. and Wienrich, C. (2022) Congruence and Plausibility, Not Presence: Pivotal Conditions for XR Experiences and Effects, a Novel Approach. *Frontiers in Virtual Reality*, **3**, Article 694433. https://doi.org/10.3389/frvir.2022.694433

[64] Lazarus, R.S., Speisman, J.C. and Mordkoff, A.M. (1963) The Relationship between Autonomic Indicators of Psychological Stress: Heart Rate and Skin Conductance. *Psychosomatic Medicine*, **25**, 19-30. https://doi.org/10.1097/00006842-196301000-00004

[65] Zhang, Z.H. and Fort, J.M. (2023) Facial Expression Recognition in Virtual Reality Environments: Challenges and Opportunities. *Frontiers in Psychology*, **14**, Article 1280136. https://doi.org/10.3389/fpsyg.2023.1280136

[66] Zhao, Y.H. and Shu, X.Q. (2023) Speech Emotion Analysis Using Convolutional Neural Network (CNN) and γ Classifier-Based Error Correcting Output Codes (ECOC). *Scientific Reports*, **13**, Article No. 20398. https://doi.org/10.1038/s41598-023-47118-4

[67] McLennan, S., *et al.* (2022) Embedded Ethics: A Proposal for Integrating Ethics into the Development of Medical AI. *BMC Medical Ethics*, **23**, Article No. 6. https://doi.org/10.1186/s12910-022-00746-3