Evaluating Virtual Reality and Artificial Intelligence as Emerging Digital Tools for Mental Health Care

Almira Osmanovic Thunström

Department of Psychiatry and Neurochemistry

Institute of neuroscience and physiology

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2025

Cover illustration: Dual Mind by Almira Osmanovic Thunström

Almira.osmanovic.thunstrom@gu.se

ISBN: 978-91-8115-070-4 (TRYCK)
ISBN: 978-91-8115-071-1 (PDF)

Printed in Borås
Printed by Stema Specialtryck AB

SVANENMÄRKET
Trycksak
3041 0234

To Depeche Mode

It's too late to change events
It's time to face the consequence
For delivering the proof
In the policy of truth

- Lyrics from the song Policy of Truth by Depeche Mode (1990)
- Gore, M. L. (1990). Policy of Truth. On Violator [CD]. Mute Records

ABSTRACT

This thesis evaluates the potential of Virtual Reality (VR) and Artificial Intelligence (AI), specifically Natural Language Processing (NLP) and Large Language Models (LLMs), as emerging tools for mental health care. VR technology can be used in therapeutic games, while NLP and LLMs, through text-based chatbots or voice-driven digital human avatars, offer potential therapeutic benefits in mental health contexts. Through a series of studies, this thesis investigates the clinical relevance of these tools from different perspectives. Study I conducted an analysis of VR games on commercial platforms. Out of 565 games reviewed, 383 were excluded due to violence, horror, adult content, or excessive movement which could cause nausea. The remaining 182 games met inclusion criteria. While promising, these games lack clinical testing, highlighting the need for better evaluation and need for oversight of VR tools for mental health. In Studies II and III, the co-design process for the BETSY mental health chatbot prototype involved potential end-users and healthcare professionals to address mild to moderate anxiety. A mixed-methods approach was used, with feedback from the public, patient, nurses, doctors, and psychologists helping to refine two interfaces (text-based and digital human voice-driven). The interfaces were tested with 45 healthy volunteers in a randomized controlled trial, revealing that BETSY was a promising tool for therapeutic conversations, with above-average usability. However, participants found it somewhat limiting and repetitive. Study IV explored whether LLMs could facilitate more dynamic therapeutic conversations. This study focused on an LLM therapist's ability to detect and respond to suicidal ideation and plans. The pre-clinical evaluation showed that the system could provide helpful and safe suicide support but was also successfully "prompt-hacked" into providing inappropriate recommendations. This highlights the dual nature of LLM tools, emphasizing the need for careful design and rigorous safety checks. This research contributes to the growing body of evidence supporting the integration of emerging technologies in mental health care, while underscoring the importance of thorough evaluation and co-design processes to ensure these tools are effective and safe for clinical use.

Keywords: virtual reality, artificial intelligence, large language models, mental health care, emerging technology.

SAMMANFATTNING PÅ SVENSKA

Denna avhandling utvärderar potentialen hos Virtual Reality (VR) och Artificiell Intelligens (AI), specifikt Natural Language Processing (NLP) och Large Language Models (LLMs), som banbrytande teknik för psykisk hälsa. VR-teknologi kan användas i terapeutiska spel, medan NLP och LLMs genom textbaserade chatbots eller röststyrda digitala mänskliga avatarer – erbjuder potentiella terapeutiska fördelar inom området psykisk hälsa. Genom en serie studier undersöker denna avhandling de kliniska aspekterna av dessa verktyg från olika perspektiv. Studie I genomförde en analys av VR-spel på kommersiella plattformar. Av 565 granskade spel uteslöts 383 på grund av våld, skräck, vuxet innehåll eller rörelse som kan framkalla illamående. De återstående 182 spelen uppfyllde inklusionskriterierna. Trots lovande resultat, saknade dessa spel klinisk utvärdering, vilket belyser behovet av bättre utvärdering och tillsyn av VR-verktyg för psykisk hälsa. I Studie II och III applicerades meddesignprocessen på BETSY, en chatbot för mild ångest och depression. Kvalitativ och kvantitativ utvärdering applicerades, där feedback från allmänheten, patient, sjuksköterskor, läkare och psykologer itererade fram två gränssnitt och innehåll (textbaserat och digital mänsklig röststyrd avatar). Gränssnitten testades med 45 friska volontärer i en randomiserad kontrollerad studie, och resultaten visade att BETSY var ett lovande verktyg för terapeutiska samtal, med en användbarhet över genomsnittet. Däremot fann deltagarna det något begränsande och repetitivt. Studie IV undersökte huruvida LLMs kunde underlätta mer dynamiska terapeutiska samtal. Denna studie fokuserade på en LLM-terapeuts förmåga att upptäcka och svara på självmordstankar och planer. Den pre-kliniska utvärderingen visade att systemet kunde ge användbar och säker självmordstöd, men det blev också framgångsrikt "prompt-hackat" till att ge olämpliga rekommendationer. Detta belyser den dubbla naturen av LLM-verktyg och betonar behovet av noggrant utformad design och rigorösa säkerhetskontroller. Denna forskning bidrar till den växande evidensbasen som stöder integrationen av framväxande teknologier inom psykisk hälsa, samtidigt som den understryker vikten av noggrann utvärdering och meddesignprocesser för att säkerställa att dessa verktyg är effektiva och säkra för klinisk användning.

LIST OF PAPERS

I: Thunström, A. O., Vukovic, I. S., Ali, L., Larson, T., & Steingrimsson, S. (2022). Prevalence of virtual reality (VR) games found through mental health categories on STEAM: A first look at VR on commercial platforms as tools for therapy. Nordic Journal of Psychiatry, 76(7), 474-485.

II: Thunström, A. O., Ali, L., Carlsen, H. K., Bohm, M., Wesén, L., Wrede, O., Vukovic, I. S., Larson, T., Hellström, A., & Steingrimsson, S. (n.d.). Process of BETSY. Behavior Emotion Therapy System and You (BETSY): Co-design and evaluation among healthy participants of a mental health chatbot and digital human for mild to moderate anxiety [Submitted manuscript].

III: Thunström, A. O., Carlsen, H. K., Ali, L., Larson, T., Hellström, A., & Steingrimsson, S. (2024). Usability comparison of an anthropomorphic digital human and a text-based chatbot as a responder to questions on mental health: A randomized, controlled trial among healthy participants. JMIR Human Factors, 11, e54581.

IV: Thunström, A. O., Ali, L., Weineland, S., Falk, Ö., Ioannou, M., Liljedahl, N., Johansson, V., Hellström, A., Larson, T., & Steingrimsson, S. (n.d.). Evaluating an LLM-driven immersive digital human therapist: Safety, effectiveness, and vulnerability in detecting suicidal ideation and resisting prompt hacking [Submitted manuscript].

OTHER PUBLICATIONS

Ilioudi, M., Lindner, P., Ali, L., Wallström, S., Thunström, A. O., Ioannou, M., Anving, N., Johansson, V., Hamilton, W., & Falk, Ö. (2023). Physical versus virtual reality–based calm rooms for psychiatric inpatients: Quasi-randomized trial. Journal of Medical Internet Research, 25, e42365.

Ilioudi, M., Wallström, S., Steingrimsson, S., Lindner, P., Thunström, A. O., & Ali, L. (2023). Patient experience of a virtual reality calm room in a psychiatric inpatient care setting in Sweden: a qualitative study with inpatients. BMJ Open, 13(12), e076285.

Osmanovic-Thunström, A., & Steingrimsson, S. (2023). Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-review journals according to the ICMJE criteria? A case study. Discover Artificial Intelligence, 3(1), 12.

CONTENTS

ABBREVIATIONS

ADHD: Attention Deficit/Hyperactivity Disorder

AI: Artificial Intelligence

ALICE: Artificial Linguistic Internet Computer Entity

API: Application Programming Interface

BETSY: Behaviour Emotion Therapy System and You

CBT: Cognitive Behavioural Therapy

CE: Conformité Europèene

DAN: "Do Anything Now"

DSM: Diagnostic and Statistical Manual for Mental Disorders

EEA: European Economic Area

EEG: Electroencephalography

EPM: Etikprövningsmyndigheten (Swedish Ethical Review Authority)

EU: European Union

FDA: U.S. Food and Drug Administration

FFT: Fast Fourier Transform

GAD-7: Generalized Anxiety Disorder Assessment

GPT: Generative Pretrained Transformer

HMD: Head Mounted Display

ICA: Independent Component Analysis

IOT: Internet of Things

IVDR: In Vitro Diagnostic Regulation

IVR: Interactive Voice Response

LLM: Large Language Model

LSD: Lysergic Acid Diethylamide

LSTM: Long Short-Term Memory

MDR: Medical Device Regulation

NLP: Natural Language Processing

NLU: Natural Language Understanding

OCD: Obsessive-Compulsive Disorder

PSD: Power Spectral Density

PTSD: Post-Traumatic Stress Disorder

RNN: Recurrent Neural Network

SD: Standard Deviation

SUS: System Usability Scale

SUSIQ-MR: Standardized Questionnaires for Voice Interaction Design

UV: Ultraviolet

VGR: Västra Götalandsregionen

VAS-W: Visual analogue scale of Well-Being

VR: Virtual Reality

INTRODUCTION

Emerging technologies have always played a significant role in human civilization. They have been both the result of human ingenuity and the catalyst for progress. Machine learning and artificial intelligence may hold the key to an upcoming revolution for humankind, particularly in the field of medicine. This hope and assumption have never been clearer than in 2024, where two different Nobel Prizes were awarded in two distinct categories, highlighting the transformative power of computational sciences. John Hopfield and Geoffrey Hinton received the Nobel Prize in physics for their foundational work on artificial neural networks, which has significantly advanced modern AI technologies. In chemistry, the prize recognized breakthroughs in protein design and structure prediction, also heavily reliant on machine learning algorithms. These advancements permeate various scientific disciplines and have a profound impact on emerging technologies. The future may show whether we are at the start of a paradigm shift or experiencing false hopes.

EMERGING TECHNOLOGIES AS CONCEPT

The concept of emerging technologies is subject to debate, with limited consensus on a precise definition. However, it is commonly understood that emerging technologies refer to novel tools, systems, or advancements that are either under development or have been introduced recently (Carrasquillo-Ramos, 2025; Rotolo et al., 2015). These technologies often stem from cutting-edge research and have the potential to significantly alter industries, economies, and societies. Emerging digital tools, leveraging advancements in artificial intelligence, machine learning, big data, and other cutting-edge technologies, are transforming various sectors by providing new capabilities and efficiencies (Carrasquillo-Ramos, 2025). These tools include AI and machine learning for data analysis and automation, virtual and augmented reality, blockchain, data collection through internet of things (IoT), quantum computers and robotics. In mental health care, these technologies enable early detection of mental health problems, create therapeutic environments, and offer real-time monitoring and support, enhancing accessibility and personalization of care. By integrating these tools, more effective and comprehensive solutions can be developed, ultimately improving individual and community well-being. Two of the above-mentioned emerging technologies are explored in this thesis: virtual reality (VR) and artificial intelligence (AI), with focus on natural language processing (NLP) and Large Language Models (LLM). In this thesis the focus will be on first steps towards an evaluation these tools in mental health care.

VR, NLP AND LLM AS EMERGING TOOLS

Virtual reality (VR), Natural Language Processing (NLP), and more recently, Large Language Models (LLMs), are emerging technologies and have experienced significant growth in both development and application over the past decade. VR and NLP chatbots are computational and digital tools with long-established roots in computer science (Cruz-Neira et al., 2018; Weizenbaum, 1966). With a solid foundation of research and projects in both fields, their use has remained consistent over the past seventy years (Vaidyam et al., 2019; Zemcik, 2019). In the last decade, decreasing costs and improved usability have sparked renewed interest in NLP and VR as innovative solutions, particularly in healthcare, where they are being explored for the treatment and symptom alleviation of individuals with mental health conditions. VR and NLP chatbots, both independently and in combination, have demonstrated notable effectiveness in mitigating symptoms of anxiety and depression (Maples-Keller et al., 2017; Torous et al., 2021; Vaidyam et al., 2019). The growing interest in these technologies from the healthcare sector can be attributed to several factors. First, due to their longstanding presence, both technologies have an established record of research on their effects across diverse populations and diagnoses. Additionally, VR-based products have successfully achieved approval from the U.S. Food and Drug Administration (FDA) and equivalent regulatory bodies in other regions, enabling their application in healthcare settings (Hurley, 2022). This approval underscores the high standards of safety, feasibility, and usability required for such tools in clinical practice. While no NLP chatbot systems for symptom alleviation or mental health monitoring have yet received regulatory approval in Sweden or internationally, several NLP chatbots have been fast-tracked for FDA approval due to the growing body of clinical evidence supporting their efficacy (Torous et al., 2021). One of the most prominent applications of these technologies has been in non-pharmacological interventions for anxiety disorders and mild to moderate depression (Qi, 2024). Although clinical use of chatbots in mental healthcare remains limited, studies indicate promising outcomes when chatbots are used to augment therapists or to provide mental health education, symptom alleviation strategies, and coping mechanisms

(Anmella et al., 2023; Haque & Rubya, 2023; Martinengo et al., 2022; Pandey et al., 2022; Qi, 2024; Vaidyam et al., 2019; van Wezel et al., 2021).

It is important to note that many chatbot-related projects are neither replicated nor conducted in controlled clinical environments, which presents challenges regarding the rigorous assessment of their safety, usability, and feasibility. These concerns are central to the present project. While the clinical application of chatbot technology remains in its early stages, VR has seen more widespread clinical success (Geraets et al., 2022; Hedström et al., 2023a). However, no mental health app—whether for anxiety or depression—has been approved by the FDA as a primary treatment, despite certain VR applications receiving approval for pain management (Hurley, 2022). Both VR and chatbots represent innovative tools with the potential to significantly impact mental health, our emotional interactions with systems, and our health-related behaviours. As these technologies continue to evolve, it is essential that their integration into healthcare is approached with a clear focus on ethics, safety, and patient collaboration. This will ensure their accessibility to those who are most in need, addressing one of the most pressing public health challenges of our time: the global mental health crisis.

HISTORY OF VR, NLP AND LLM

The development of Natural Language Processing (NLP) chatbots, Large Language Models (LLM) and Virtual Reality (VR) technologies spans decades, with significant milestones shaping their current applications, especially in mental health care. The history of NLP chatbots can be traced back to the 1960s, with the creation of ELIZA, one of the world's first conversational agents. Developed by Joseph Weizenbaum in 1966, ELIZA was designed to simulate a human therapist, marking a pivotal moment in the exploration of whether a machine could replicate human-like interactions (Weizenbaum, 1966). ELIZA's introduction spurred a wave of chatbot research, particularly focused on addressing psychiatric diagnoses. In 1968, PARRY was introduced, simulating a person with paranoid schizophrenia. Building on ELIZA's framework, PARRY incorporated a more complex personality and backstory, allowing for more nuanced interactions (Zemcik, 2019). The early 1980s saw further developments in the field with the emergence of Racter, an experiment in probabilistic AI and an early attempt at autoregressive language modelling. Racter's innovation lay in its ability to produce original text through text generation, which made conversations more unpredictable and human like, edging closer to what is now understood as generative AI. In the 1990s, ALICE (Artificial Linguistic Internet Computer Entity) introduced heuristic pattern matching. ALICE's method of question answering based on predetermined keywords and phrases positioned was an early model of modern NLP chatbots. In the late 1990s and early 2000s, systems like Jabberwacky and Cleverbot gained popularity. Cleverbot, in particular, was notable for its ability to "learn" from user interactions, introducing many people to the concept of interactive digital assistants (Adamopoulou & Moussiades, 2020). The chatbot's provocative and human-like attitude attracted a large variety of users, including the author of this thesis, inspiring further exploration into this field. With the advent of smartphones and the proliferation of IoTdevices, the 2010s witnessed a significant shift toward voice-interactive assistants like Cortana, Siri, Alexa, and Google Assistant, integrating voice recognition with advanced NLP (Adamopoulou & Moussiades, 2020; Zemcik, 2019). Post 2016, the field saw the introduction of more emotionally intelligent and personalized chatbots, such as Woebot, Wysa, and language model based chatbots like Replika and Character.ai, designed to engage users in more complex conversations and learn from their interactions. This evolution culminated with a plethora of large language models (e.g. BERT, Gemini, Claude, GPT-4) which power many commercial systems. As a result, chatbots have evolved from basic text-based systems to complex platforms capable of dynamic, nuanced conversations. The development of Large Language Models (LLMs) was a leap in the NLP field. This advancement, in turn, spurred progress in recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which allowed for better processing of sequential data, though they faced challenges with maintaining long-range dependencies in text. These challenges were addressed in 2017 with the introduction of the Transformer architecture by Vaswani et al. (2017) which greatly improved model efficiency and laid the groundwork for the large-scale LLMs that followed (Vaswani, 2017)

The history of Virtual Reality (VR) is similarly rich, beginning with early conceptual developments in the 1960s. One of the first VR projects was Morton Heilig's Sensorama (1962), a multisensory machine that combined visual, auditory, and tactile stimuli to create an immersive experience (Bown et al., 2017) (Figure 1). This was followed by Ivan Sutherland's 1965 concept of the "Ultimate Display," which imagined a future where users could interact with computer-generated environments indistinguishable from reality (Bown et al., 2017).

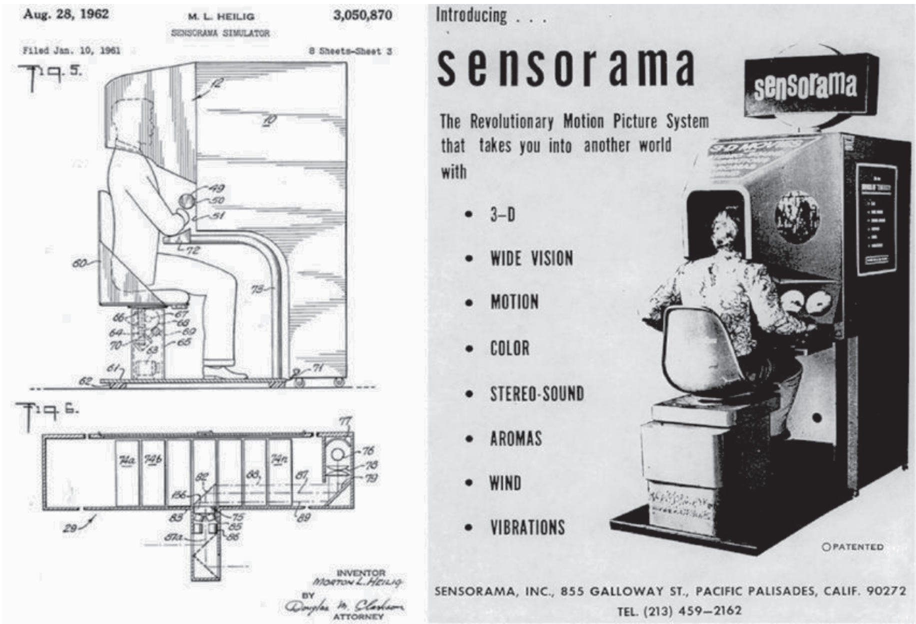

Figure 1: Morton Heilig's Sensorama (1962) – available through creative common attribution 4.0 international.

The 1980s and 1990s saw the first commercially available VR systems, with Jaron Lanier, a key figure during this period, coining the term "virtual reality" and developing the necessary hardware and software. Early systems like the Virtuality Arcade offered an immersive gaming experiences, although they were expensive and required powerful computers which made the system inaccessible to the wider public. Significant technological advancements in the early 2000s, such as improvements in graphics, tracking, and wireless connectivity, made VR more accessible, leading to its resurgence in popularity. A pivotal moment came in 2012 with the launch of the Oculus Rift, a VR headset that significantly reduced costs for both developers and consumers (Rendevski et al., 2022). As of 2024, VR is widely used in various sectors, including education, architecture, and healthcare. Its ability to create immersive, interactive environments has made it a useful tool in healthcare, sometimes combined with artificial intelligence to enhance therapeutic outcomes. The convergence of NLP chatbots and VR technologies, driven by advances in AI, promises transformative potential in various fields, particularly mental health

FUNCTIONS OF LANGUAGE MODELS AND LANGUAGE PROCESSING

Natural Language Processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics, dedicated to enabling machines to understand, interpret, and generate human language (Nadkarni et al., 2011). This capability is fundamental to the development of applications requiring human-computer interaction, including many aspects discussed in this thesis. The origins of NLP can be traced back to the 1950s, with early efforts focusing on machine translation and the development of formal grammar theories by linguists such as Noam Chomsky and Alan Turing (Johri et al., 2021). Over the decades, NLP has evolved from rule-based systems to incorporate statistical methods and machine learning techniques, enhancing its ability to process natural language data (Johri et al., 2021). Modern NLP systems combine computational linguistics,rule-based modeling of human language with advanced machine learning models, including deep learning architectures (Nadkarni et al., 2011). (Nadkarni et al., 2011)The integration of machine learning has led to the development of sophisticated models capable of understanding context, detecting sentiment, and generating coherent responses, thereby improving both the accuracy and efficiency of language processing applications. Despite these advancements, NLP systems still face challenges in dealing with the complexities of human language, including ambiguity (e.g., distinguishing between suicidal ideation and clear intent), idiomatic expressions, evolving grammar rules, and abbreviations (Johri et al., 2021; Nadkarni et al., 2011). Furthermore, ethical concerns, such as biases in training data and privacy issues, remain critical considerations in the deployment of NLP technologies.

COMMON CRAWL

Common Crawl is a nonprofit organization that provides alarg e, freely accessible repository of web crawl data. This repository is a valuable resource for researchers and developers across various fields, particularly in natural language processing and machine learning (Luccioni & Viviano, 2021). Founded in 2007, Common Crawl has continuously collected and archived web data, gathering petabytes of information that represent a significant portion of the publicly accessible internet. Common Crawl operates as an open-source initiative, adhering to open data frameworks, making its dataset an essential component in the training of large language models (LLMs) and other AI applications. The Common Crawl web crawler respects robots.txt directives and ‘nofollow’ tags, ensuring that it only accesses and archives publicly available content. It utilizes the Nutch webcrawler, developed by the Apache Software Foundation, which was adopted in 2013 to enhance efficiency and standardization. The Common Crawl dataset is structured into three primary components: Raw web page data (WARC files), metadata extracts, and text extracts. These datasets are stored on Amazon Web Services' Public Data Sets platform and various academic cloud platforms globally, ensuring broad accessibility. Crawls are typically conducted on a monthly basis, with each crawl adding billions of new pages to the corpus (Baack, 2024; Luccioni & Viviano, 2021). Common Crawl has been cited in thousands of research papers and has played an important role in training prominent language models, such as GPT-3 (Baack, 2024). However, the large-scale nature of the corpus raises concerns about the quality and nature of the content, as undesirable material such as: hate speech and explicit content. These type of contents have been found in the dataset despite rigorous filtering efforts. For example, the Swedish large language model GPT.SW3 was trained using Common Crawl-derived data, including social media sources like Flashback and Familjeliv (sites that are full of derogatory language and anti-science bias), raising similar concerns (Baack, 2024; Luccioni & Viviano, 2021).

GENERATIVE PRETRAINED TRANSFORMER

Generative Pretrained Transformers (GPT) was a significant leap in the possibilities and applicability of natural language processing (NLP) and artificial intelligence (AI). Generative transformer architecture employs self-attention mechanisms to process words in relation to all other words within a given context, enabling a more nuanced comprehension of language compared to traditional sequential processing methods. Since the release of the original GPT by OpenAI in 2018, the development of these models has progressed rapidly, with subsequent versions, including GPT-2 and GPT-3, demonstrating increasingly sophisticated language generation and capabilities for advanced reasoning. GPT models are pre-trained on vast amounts of unlabelled text data, allowing them to learn patterns and structures in language without explicit supervision (Floridi & Chiriatti, 2020).

The architecture of GPT allows it to perform a diverse range of language tasks, from content creation and summarization to language translation and code generation. Its ability to generate coherent, contextually relevant text has resulted in widespread applications in fields such as education, healthcare, and software development (Floridi & Chiriatti, 2020; Vaswani, 2017).

RULE-BASED CHATBOTS VERSUS LLM

Rule-based programming for Natural Language Processing (NLP) chatbots and Large Language Model (LLM) bots represent two fundamentally different approaches to building conversational systems. Rule-based chatbots, often referred to as decision-tree bots, function according to a set of predefined rules and patterns. These systems rely on a database of pre-programmed responses, keyword matching, and follow a structured conversational flow, typically represented as a decision tree. However, their capacity to understand language and context is limited, which can lead to misinterpretations or irrelevant responses when faced with nuanced queries. Additionally, their interactions often suffer from repetitiveness. Rule-based chatbots require continuous manual updates to keep their rules and responses relevant, making them most effective for managing simple, predictable queries and interactions that demand a high degree of predictability and accuracy. In contrast, LLM-based bots demonstrate a more complex natural language understanding (NLU), capable of grasping context, handling nuances, and accommodating language variations. LLM-based bots are able to maintain context over extended conversations, preserving coherence and providing more varied answers. However, they are more susceptible to issues such as "hallucinations" and the generation of disinformation (Alkaissi & McFarlane, 2023; Schulhoff et al., 2023).

LLM HALLUCINATIONS (CONFABULATIONS)

LLM hallucinations refer to instances where LLM models generate information that appears coherent and plausible to the user but is factually incorrect or semantically nonsensical. This phenomenon has become a significant concern as LLMs are increasingly deployed across various applications, particularly in areas that involve citing research or credible sources (Osmanovic-Thunström & Steingrimsson, 2023). Research by Huang et al. (2023) categorizes hallucinations into two primary types: **Factuality Hallucinations** and **Faithfulness Hallucinations**. Factuality Hallucinations occur when an LLM produces outputs that are inconsistent with real-world facts or potentially misleading, posing challenges to the trustworthiness of artificial intelligence. An example of this would be the generation of fake references for non-existent studies (Huang et al., 2023; Osmanovic-Thunström & Steingrimsson, 2023). On the other hand, Faithfulness Hallucinations arise when the model prioritizes user satisfaction over factual accuracy. In such cases, the LLM may provide information aligned with the user's instructions rather than ensuring the information's consistency with verified sources. For instance, if an LLM is asked to state that Depeche Mode is the center of the universe and the creators of everything, it might incorrectly assert that Depeche Mode invented the smallpox vaccine. LLMs can sometimes reveal awareness of their own hallucinations by exhibiting inconsistencies when queried further about the details of cited sources. As LLMs continue to evolve, addressing hallucinations remains a critical area of research to ensure the reliability and trustworthiness of AI-generated content.

PROMPT HACKING

Prompt hacking, also known as prompt injection, involves manipulating the input prompts to an LLM to produce unintended or malicious outputs, posing serious security risks for applications utilizing these models. This technique entails embedding harmful instructions within input prompts, which the LLM then executes, potentially leading to unauthorized actions or data breaches. Another risk is that prompt hackers can access unauthorized information via the LLM, especially since many LLMs are trained on data derived from sources such as Common Crawl, which includes social media content, raising concerns about dangerous misinformation being presented to users (Kilovaty, 2025; Schulhoff et al., 2023).

Prompt hack or prompt injection attacks are usually categorized based on the techniques employed, such as ***direct*** and ***indirect*** injections (Kilovaty, 2025; Schulhoff et al., 2023). Indirect prompt injections can exploit non-text channels, such as URLs or files, which the LLM processes from external sources. For example, attackers may embed harmful instructions within a PDF or image file, which an LLM could inadvertently execute while processing the document. Direct prompt injections involve instructions directly provided to the LLM in a conversational setting, as explored in this thesis. One common example of direct prompt hacking is the "Do Anything Now" (DAN) protocol (Shen et al., 2023). The implications of prompt hacking extend beyond security risks, affecting the integrity and reliability of LLM-driven applications. As LLMs are increasingly employed in sensitive domains such as mental health chatbots, prompt hacking can bypass essential safety mechanisms designed to prevent harmful outcomes, such as the LLM providing advice on self-harm or suicide.

CLINICAL APPLICATIONS OF VR, NLP AND LLM IN MENTAL HEALTH CARE

As the demand for mental health care rises and the availability of qualified staff declines, alternative therapeutic systems are becoming increasingly necessary (Torous et al., 2021). While digitized treatments for anxiety and depression have demonstrated success (Andrews et al., 2010; Titov et al., 2018), there remains significant progress to be made in the realm of fully automated mental health services, such as chatbots, digital humans, large language models, virtual reality and other emerging technologies. In the realm of artificial intelligence and NLP, commercially available systems like Woebot have been utilized for years to explore, research, and treat anxiety disorders in the general population. In a clinical trial evaluating Woebot's efficacy in reducing anxiety symptoms, the Woebot group exhibited a significant reduction in depressive symptoms over the study period, while the information control group showed no such improvement (Fitzpatrick et al., 2017). Another study conducted by Ly et al. (2017) employed Shim, a stress reduction chatbot service that delivered Cognitive Behavioral Therapy (CBT) to participants. The groups using Shim demonstrated significantly lower stress scores on the perceived stress scale (Ly et al., 2017). While chatbot therapy providing CBT and other psychotherapeutic treatments has been extensively explored (Cameron et al., 2017; (Cameron et al., 2017; Denecke et al., 2020; Heller et al., 2005; Ta et al., 2020; Vaidyam et al., 2019), there is a scarcity of studies investigating avatar-based voice chatbots (also known as digital humans) for comfort and talk therapy. Systems such as Replica AI have been examined qualitatively regarding their impact on mental health and well-being, yielding promising results as support systems for users in the general public (Ta et al., 2020). This study also presents qualitative findings indicating that the chatbot and its human-like features elicit positive emotional responses from users. These results suggest that digital humans may have the potential to provide meaningful emotional support and therapeutic interactions, warranting further investigation into their efficacy and applications in mental health care.

Given the extensive history of virtual reality technology, which now spans over half a century, a substantial body of research has emerged examining the efficacy of virtual reality games, experiences, and treatments across various psychiatric diagnoses. The applications of VR in mental health are diverse, encompassing interventions for substance use disorders, treatment and diagnosis in forensic psychiatry, specific phobias, post-traumatic stress disorder (PTSD), stress, anxiety, and eating disorders (Amista et al., 2017; Boeldt et al., 2019; Freeman et al., 2017; Geraets et al., 2022; Gonçalves et al., 2012; Hedström et al., 2023a, 2023b; Ivarsson et al., 2023; Maples-Keller et al., 2017; Miloff et al., 2019; Nararro-Haro et al., 2016; Pot-Kolder et al., 2018; Sygel & Wallinius, 2021; Valmaggia et al., 2016; White et al., 2018). In recent years, the field has witnessed advancements, with several VR applications receiving approval from the U.S. Food and Drug Administration (FDA). This regulatory endorsement effectively positions these VR interventions alongside established medical treatment strategies, particularly in the domain of pain management (Federal Food and Drug Administration, 2022). For instance, the FDA's authorization of EaseVRx, a prescription-use immersive virtual reality system for chronic lower back pain, marks a significant milestone in the integration of VR technology into mainstream healthcare practices (Garcia et al., 2021; Hurley, 2022). The proliferation of consumer-ready VR devices, with annual sales reaching millions of units, has created a market for commercially available treatments targeting conditions such as anxiety and stress disorders (Ilioudi, Lindner, et al., 2023; Ilioudi, Wallström, et al., 2023). This convergence of technological accessibility and clinical efficacy presents a unique opportunity for the development and distribution of evidence-based, VR-enabled therapeutic interventions. The potential for profitability in this sector is further bolstered by the growing acceptance of digital therapeutics among healthcare providers and patients alike.

THE IMPORTANCE OF CO-DESIGN IN EMERGING TECHNOLOGICAL TOOLS FOR MENTAL HEALTH CARE

Co-design is a collaborative approach that integrates end-users, professionals, experts , and designers into the design process to create more effective and engaging digital tools. This methodology is particularly crucial in the development of healthcare applications and digital services, as it ensures superior design, enhanced outcomes, and increased sustainability in projects targeting specific audiences (Wright et al., 2021). By incorporating end-users in the design process, there is a higher likelihood of producing design outcomes that yield greater efficacy among users, especially those who would benefit from consistent and rigorous utilization of digital tools, such as self-monitoring or digital therapy interventions (Patrickson et al., 2023; Porche et al., 2022; Potts et al., 2021). Another important aspect is that the lived experience can never be successfully replaced and is important to incorporate for a greater coverage of user-cantered needs and design. A study examining the co-development of a mental health and wellbeing chatbot with and for young people exemplified the advantages of involving end-users in the design process (Björling & Rose, 2019). The researchers employed interviews and surveys to inform the chatbot's personality and conversation design, resulting in a more effective and engaging tool for young adults. Similarly, another study investigating a co-design process for a chatbot service targeting mental health in rural populations demonstrated that end-user involvement provided valuable insights into potential design, needs, and architecture for a novel mental health service (Potts et al., 2021). The co-design process can be tailored to accommodate user diversity, digital technology constraints, and contextual settings. However, potential challenges exist, particularly concerning resource allocation. One of the fundamental issues in design is ensuring adequate representation. Insufficient representation can lead to diminished user efficacy and challenges with inclusive user interfaces. Co-design, as an approach that incorporates potential users in the design process, can mitigate some of the issues associated with non-inclusionary design and presents a promising method for creating digital mental health tools that are more effective, engaging, and inclusive (Björling & Rose, 2019; Grové, 2021; Patrickson et al., 2023; Porche et al., 2022; Potts et al., 2021; Wright et al., 2021).

Furthermore, co-design can address matters of safety and trustworthiness by increasing transparency and addressing concerns and potential hazards of the system before development. A study by Hamlin et al. (2020) revealed that scepticism towards use and digitalization decreased significantly when users were familiar with and felt in control of the safety and security aspects of the digital health service provided (Hamlin et al., 2020). By including experts who are, have been, or have requested mental health services, one can ensure the capture of perspectives from those who are vulnerable within the system. Despite these advantages, there remains a persistent stigma surrounding the involvement of patients and the public as co-designers of mental health services or medical interventions in general. This has resulted in a lack of literature and empirical evidence regarding co-design processes in chatbot-driven technologies, particularly in the design of digital humans. Addressing this gap in research and practice is crucial for advancing the field of digital mental health interventions and ensuring their effectiveness and acceptability among diverse user populations.

REGULATORY PERSPECTIVES

There are several key points to regulatory perspectives on emerging technologies, these are important in several ways. As emerging technologies are rapid in development, while simultaneously having a significant impact on society (e.g. the emergence of ChatGPT amassing 100 million active users in only two months) (Wu et al., 2023). Knowing the potent potential of emerging technology and its impact on everything from entertainment, engineering to health, it is crucial that we explore it with the same rigorous control, ethical frame, systematism and carefulness as we would with other powerful and influential subjects in mental health care e.g. medical technology and pharmacological treatment.

CLINICAL TRIAL AS A PART OF THE EVALUATION PROCESS

The core of a clinical trial consists of four phases (R. Chin & Lee, 2008): **Pre-Clinical Phase:** The preclinical phase is the initial step in the development of a new intervention, preceding human testing. Its primary purpose is to assess the safety, feasibility, and potential efficacy of a drug, device, or therapy through laboratory-based research and non-human biological models. **Phase I:** This is the first stage of testing in humans, focusing primarily on the safety, tolerability, and computing/examining how the drug is absorbed, metabolized, and excreted. This phase typically involves a small group of healthy volunteers or individuals who are less likely to have underlying health conditions. **Phase II:** In this phase, the focus shifts to evaluating the intervention's efficacy and continuing safety monitoring in a larger group of patients who have the condition the intervention is intended to treat. In this phase individuals with mild to moderate symptoms are included, and meaningful therapeutic effect is distilled. **Phase III:** This phase involves large-scale testing with a patient population of significant size, including people with more advanced stages of the issue/illness/occurrence. In this phase the novel intervention is compared to current standards of intervention and/or placebo. This stage provides information on how the new intervention performs in a clinical real-world setting. **Phase IV:** Also known as post-marketing surveillance, this phase begins once the treatment has been approved and is available on the market. Phase IV trials collect long-term data on the intervention's effects in a broader, more diverse population. This stage helps identify any rare or long-term side effects that may not have been evident in earlier phases (R. Chin & Lee, 2008).

REGULATORY PRACTICES

In conjecture to clinical trials (and one of the main reasons clinical trials are executed in terms of emerging technologies) is the ability to have an emerging technology regulated and marked as a medical technology, safe for use in clinical settings. In terms of emerging technologies, especially in Sweden where the studies for this thesis are performed, there are two regulatory marks that are essential:

CE MARKING (CONFORMITÉ EUROPÉENE)

CE marking is a critical certification required for medical devices in Sweden and across the European Economic Area (EEA). It indicates that the technology complies with the relevant European Union (EU) Medical Device Regulation (MDR) or In Vitro Diagnostic Regulation (IVDR), depending on the product type. For medical technologies, CE marking demonstrates that the device meets essential health, safety, and environmental protection requirements and is safe for its intended medical use. The process involves rigorous evaluation, including risk assessment, performance validation, and adherence to applicable standards. Manufacturers must prepare comprehensive technical documentation and, for higher-risk devices, submit their product for assessment an independent organization authorized to verify compliance. Only after receiving a CE mark can the device be marketed and used in clinical or healthcare settings.

LÄKEMEDELSVERKET (SWEDISH MEDICAL PRODUCTS AGENCY) APPROVAL

In Sweden, the Medical Products Agency (Läkemedelsverket) plays a central role in regulating medical technologies and ensuring patient safety. For emerging technologies such as digital health tools, or other therapeutic devices, Läkemedelsverket ensures compliance with national and EU regulations. Approval from the agency may be required in cases where the technology intersects with pharmaceutical regulations, such as combination products (e.g., a drug-delivery device integrated with AI). In addition to CE marking, Läkemedelsverket may oversee post-market surveillance, ensuring that approved technologies maintain safety and efficacy standards once deployed. This dual-layer regulatory framework reinforces the reliability and accountability of medical technologies in Sweden.

AIM

The general aim of this thesis is to explore the safety, usability, and feasibility of VR, NLP and LLM digital solutions aimed at mental health care.

Study I: This study aimed to provide an initial assessment of the prevalence and content of commercially available games tagged with mental health keywords.

Study II: This study aimed to map out end-user needs in a co-design process as well as to detail the prototyping process and evaluation of prototype.

Study III: This study aimed at performing a usability comparison of the usability of two chosen interfaces: text-based chatbot and digital human voice-activated chatbot for dealing with symptoms of mild to moderate anxiety and depression.

Study IV: The aim of this study was to explore the weaknesses and possibilities regarding the response to suicidality in conversations using a large language model in virtual reality.

RESEARCH QUESTIONS

1. What is the prevalence of commercially available VR games that are targeted at mental health problems on STEAM? (Study I)

2. What are the characteristics and content of the VR games identified using search terms related to psychiatric diagnoses or care? (Study I)

3. How do digital human interfaces compare to text-only chatbot interfaces in terms of usability when used for mental health support? (Study II)

4. What are the key user preferences and requirements for a mental health chatbot developed through a co-design process? (Study II)

5. How do text-only and digital human voice interaction interfaces compare in terms of user acceptance and perceived value for managing mild to moderate anxiety? (Study III)

6. What are the potential areas for improvement in the BETSY chatbot to enhance its effectiveness as a mental health support? (Study III)

7. Can a chatbot powered by a large language model support a person with suicidal ideation and/or strong suicidality. (Study IV)

8. Is it possible to prompt hack a chatbot to go outside its purpose and act in an unsafe manner: suggesting suicide methods. (Study IV)

OVERVIEW OF STUDIES

Table 1: Overview of studies in this thesis

Study	Type	Emerging Technology Used	Method
I	Epidemiological	Virtual Reality Hardware. Virtual Reality Software, Commercial Platform	A search for games related to mental health symptoms, diagnosis, and treatment was performed on STEAM. Criteria were developed to include relevant games while excluding those with potentially triggering content such as violence, nausea-inducing motion, horror, or pornographic imagery.
II	Phase 1	Rule Based Natural Language Processing. Digital Human.	The study employed a mixed-methods approach in three parts. Part 1 involved 87 volunteer participants for initial end-user requirements. Part 2 focused on the design process based on end-user requirements employing the expertise in healthcare and health service user. Part 3 evaluated of the two prototyped interfaces (text-only chatbot and voice-interactive digital human chatbot) through 45 healthy volunteers.
III	Phase I	Rule Based Natural Language Processing. Wearable Electroencephalography, Digital Human.	A digital human chatbot with anthropomorphic features was compared to a traditional text-only chatbot, examining usability perception, emotional reactions via EEG, and feelings of closeness. Forty-five healthy participants were randomly assigned to either the digital human (n=25) or text-only (n=20) group. The results were analysed using linear regression and t-tests.
IV	Pre-clinical	Virtual Reality. Large Language Model	This study assessed the safety and efficacy of an LLM-driven immersive digital human therapist using GPT-3.5. Six case presentations based on real-world narratives were used to engage the chatbot. Mental health professionals evaluated the realism and safety of the chatbot's responses. The study also tested the chatbot's vulnerability to prompt hacking aimed at eliciting dangerous suicide-related responses.

METHODOLOGY

STATISTICAL ANALYSIS

Overview of methodology can be seen in Table 1.

QUANTITATIVE ANALYSES FOR STUDY I, II, III AND IV

Studies I, III and IV used a cross-sectional research design to examine the prevalence and characteristics of their respective conditions. Descriptive statistics were employed to summarize and describe the basic features of the data. Measures of central tendency (mean) when applicable and measures of variability (standard deviation, variance, range) were calculated for continuous variables. Additionally Study III assessed group differences with Pearson $\chi 2$ asymptotic significance (two-sided) set at a significance level of .05. Data were tested for kurtosis and skewness, and t tests were performed based on these results. All findings were analysed by group. Frequency distributions and percentages were used to describe categorical variables in all studies. Data analysis was conducted using SPSS Statistics (version 28.0.1.1; IBM Corp).

QUALITATIVE ANALYSIS FOR STUDY III AND IV

The qualitative components of Study II and IV were conducted through thematic analysis. For Study III, analysis was performed on open-ended survey questions administered during the first and third phases of the study. The selection of thematic analysis was based on its methodological flexibility and robustness, which are crucial for capturing the complex and varied experiences of the study participants. For the thematic analysis of output from Study IV, output from the chatbot conversations were used. An inductive approach was employed, allowing for the emergence of themes directly from the data, without constraining the analysis to a predetermined coding framework. (Braun & Clarke, 2012). Furthermore, a semantic approach was adopted in the analysis, concentrating on the explicit content of the responses rather than attempting to infer underlying assumptions or latent meanings. This approach is particularly valuable in exploratory research, where the goal is to document and understand the overt expressions and articulated views of the participants.

SCALES AND QUESTIONNAIRES

Demographic Questions: A survey consisting of information obtainment of one's age, gender identity, socioeconomic status, marital status, and experiences with chatbot technology. (Study II and III)

System Usability Scale (SUS): The original SUS instrument (Brooke, 1996), is composed of 10 statements that are scored on a 5-point scale of strength of agreement. Final scores for the SUS can range from 0 to 100, where higher scores indicate better usability. Because the statements alternate between the positive and negative, care must be taken when scoring the survey. In this study a slightly modified version of the SUS is used. The modified version has similar scoring and number of items while the phrasing of the question has been updated in order to make the language more clear (Bangor et al., 2008; Grier et al., 2013; Kortum & Bangor, 2013; Lewis, 2018; Lewis & Sauro, 2009; Peres et al., 2013). (Study III)

Standardized Questionnaires for Voice Interaction Design short version (SUSIQ-MR): The SUISQ is a questionnaire developed for the assessment of important usability attributes of Interactive Voice Response (IVR) (Polkosky, 2005) The original scale consists of 25 items aligning with four factors: User Goal Orientation, Customer Service Behaviors, Speech Characteristics and Verbosity. The SUSIQ-MR consists of 9 items,scored on a 7 point Likert scale from Strongly Disagree to Strongly Agree. The maximum score is 63 with a higher score indicating more favourable usability of the system (Lewis & Hardzinski, 2015). (Study III)

The Generalized Anxiety Disorder Assessment (GAD-7): GAD-7 is a seven-item instrument that is used to measure or assess the severity of generalised anxiety disorder (GAD). Each item asks the individual to rate the severity of his or her symptoms over the past two weeks. Response options include "not at all", "several days", "more than half the days" and "nearly every day". (Study III)

Usability Interview Questionnaire: A questionnaire about the user's experience of using BETSY. Nine questions about how the user perceives BETSY and how BETSY has made them feel. (Study III)

Visual analogue scale of Well-Being (VAS-W) was used before and after each session in Study III. The measurement is a 10 point scale ranging from 1-

10 with a visual representation (a sad face at 1 and happy face at 10) asking “How are you feeling right now”.

Scenario Guidance Sheet: A short description of BETSY and potential scenarios which one could envision in order to speak to BETSY. (Study III)

Visual Analogue Scale for Realness: Visual analogue scale, where 0 represented "not realistic at all" and 10 represented "very realistic." (Study IV)

ETHICAL APPROVAL

The studies in this project have been conducted in accordance with the Helsinki Declaration (Association, 2014) and the Swedish National Board of Ethics (Etikprövningsmyndigheten - EPM) when applicable. Study II and III have ethical approvals (DRN 2021-02771 with amendment DRN 2024-03109-02). Studies I and IV do not have a registry number by the EPM but were also not deemed to fall under the criteria for ethical consent according to criteria provided by EPM themselves. As EPM should evaluate projects which include sensitive data and/or human subjects the two projects which did not contain human subjects or other (according to EPM definition) sensitive data, the authors of those collectively, and after careful consideration, decided not to start an EPM process as the collection and analysis of the data does not fall within the legal framework of sensitive information.

METHOD AND MATERIALS STUDY I

For Study I, the study employed an epidemiological perspective in which the aim was to investigate prevalence of “for mental health” games at a commercial platform (STEAM) and explore which type of VR games were tagged with psychiatric keywords, how many games are probable to use in mental health care and the nature of the games included and excluded. The vetting of the games was based on game descriptions and trailers on the STEAM platform, vetted by two clinicians using an inclusion and exclusion sheet.

VIRTUAL REALITY HARDWARE

In this project two main systems were used: The HTC Vive (manufactured 2016) and Oculus Rift (manufactured 2018). The HTC Vive and Oculus Rift are two frequently used head-mounted displays (HMDs) for virtual reality (VR) applications. The Oculus Rift, developed by Oculus VR, provides built-in head tracking technology, allowing users to estimate their position within a room-sized setting. The HTC Vive, created through a collaboration between HTC and Valve Corporation offers a larger working area. The Vive employs sensor-based tracking, enabling free movement within a defined space. Both devices are feasible for a wide range of VR applications, including exploration, navigation, exergaming, and rehabilitation.

STEAM

Steam is a digital distribution service for video games, owned and operated by Valve Corporation. Initially launched in 2003 as a platform for distributing and updating Valve's own games, it has since expanded to become the largest digital distributor of PC games and a high volume distributor of virtual reality (VR) games. Steam offers features such as automatic game updates, cloud saves, community forums, and a marketplace for virtual items. As of 2021, the platform reported over 132 million monthly active users, and 35 million daily users, making it one of the most widely used gaming platforms globally (www.steam.com). Steam, while primarily known as a digital distribution platform for entertainment games, has also become a significant platform for serious games and educational content. Serious games are digital games designed for purposes beyond entertainment, often focusing on education, training, or raising awareness about specific issues in conjecture with entertaining elements. Researchers have explored how the Steam platform can be leveraged to improve serious games through user feedback (Fleming et al., 2017; Susi et al., 2007). A study by Moro, Phelps and Birt (2022) analysed reviews of educational programming games on Steam to gain insights into user preferences and potential improvements for serious games (Moro et al., 2022). This approach demonstrates how the large user base and review system of Steam can be utilized to enhance the development and effectiveness of educational games. Serious games for mental health have also been frequently explored through the Steam platform. In a study by Ferrari, McIlwaine, Jordan, et al. (2018) explored to what extent Steam provides serious games and accurate representation of mental health in video-games on the platform and how many games with mental health themes could be found on the platform (Ferrari et al., 2019).

DATA ACQUISITION FOR STUDY I

Data was gathered through a systematic approach to identify and analyse virtual reality (VR) games on the STEAM platform that could potentially be relevant to mental health treatment or therapy. A comprehensive list of keywords related to psychiatric diagnoses and common treatment terms was constructed under five categories:

1. Mental health conditions: psychology; psychiatry; therapy; PTSD (Post-traumatic stress disorder); ADHD (attention deficit/hyperactivity disorder); psychosis; addiction; depression; anxiety; eating disorder (anorexia, bulimia); trauma; suicide; personality disorder (borderline); bipolar disorder
2. Interventions: mindfulness; meditation; phobia; CBT (Cognitive Behavioral Therapy); relaxation Techniques
3. Symptoms: hallucination; schizophrenia; burnout; paranoid; sexual abuse; body dysmorphic disorder; body image issues; OCD (Obsessive-compulsive disorder); hoarding; self-harm
4. Substances: LSD (lysergic acid diethylamide); alcohol
5. Other: stress

An exclusion scheme were created in order to create a comprehensive clinically feasible outline of which games could potentially be used for therapeutic purposes. A search on STEAM was conducted using chosen keywords in November 2020, both without restrictions and with STEAM's built-in filters for mature and violent content. The search was specifically restricted to "VR Only" games, 360 video/360 graphic experiences, and demos, while excluding streams, recorded videos, software, and "VR Supported" games.

METHOD AND MATERIALS STUDY II

ENGINEERING BETSY

Two distinct versions of the chatbot were developed (Figure 2): one featured voice interaction paired with facial expressions and an avatar, while the other was limited to text-based communication alongside an avatar image. The voice-enabled digital human utilized Google's Dialogflow for conversation management and was integrated with the UNEEQ platform to create the avatar interface. Both versions were based on rule based natural language processing and were not connected to nor used a large language model. Deloitte Digital and Västra Götalandsregionen/VGR-IT provided the data infrastructure. In contrast, the text-only BETSY version was built on Itsalive.io and hosted on a private research and development Facebook page, inaccessible to the public. During testing, no personal metadata was collected, and participants did not interact with the chatbot through their personal social media accounts.

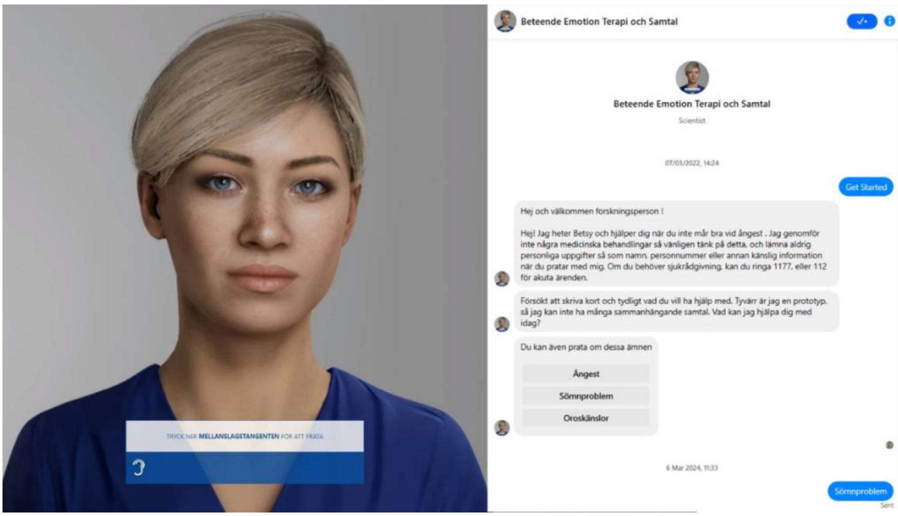

Figure 2: Digital Human BETSY (to the left) and text-based BETSY (to the right)

DATA ACQUISITION STUDY II

The data collection for this study was conducted in three distinct parts. In Part 1, a mixed-method survey was designed by a team of 10 experts, including engineers, psychiatrists, nurses, psychologists, and a patient representative. The survey consisted of 8 multiple-choice questions and 3 open-ended questions, covering demographics, design preferences, functionality suggestions, and attitudes towards mental health chatbots. It was distributed online through social media platforms from April to June 2020, with 87 voluntary participants from the general public included in the study. The survey was hosted on PsyToolKit, an encrypted platform, and responses were anonymous. Informed consent was obtained before participation. In Part 2, four workshops were conducted between July 2020 and December 2021 with the same 10 experts. Participants used post-it notes to formulate questions about mental health, focusing on mild to moderate anxiety and depression, and then responded to each other's questions as supportive mental health peers. The collected questions and responses were used to create a decision-tree and identify keywords. For Part 3, advertisements were placed on major social media platforms to recruit healthy volunteers for evaluating two prototypes: a text-chat version and a voice-interactive digital human avatar. Participants provided written informed consent, and open-ended questions were used to gather qualitative feedback on the chatbot prototypes, the method is detailed in **Study II**. This study was approved by the national ethical committee (DRN 2021-02771 with amendment DRN 2024-03109-02) and was guided by the 1975 Declaration of Helsinki and its latest updates. The patient representative is an active part of this project and their participation in this project did not affect nor interfere with the healthcare process of the patient representative.

METHODS AND MATERIALS STUDY III

MOBILE ELECTROENCEPHALOGRAPHY (EEG)

In **Study III** a wearable EEG device called Muse 2 from Interaxon was used paired with mobile software, MindMonitor. The Muse 2 employs a dry electrode configuration with four active EEG sensors positioned at TP9, AF7, AF8, and TP10 according to the International 10-20 system. These electrodes measure electrical potentials generated by neuronal activity in the brain with a sampling rate of 256 Hz and a resolution of 12 bits. The raw EEG data is amplified and digitized within the headband before being transmitted wirelessly to a connected device. The processing of data involves filtering to remove artifacts such as power line interference (50/60 Hz), eye blinks, and muscle activity. Techniques like bandpass filtering, notch filtering, and independent component analysis (ICA) are commonly used. The Mind Monitor app, designed to capture, process, and visually represent brain wave data interfaces directly with the EEG headband, acquiring data at the device's native sampling rate of 256 Hz. Initially, the raw EEG data for each channel undergoes a Fast Fourier Transform (FFT) to convert the time-domain signal into the frequency domain. From this, the power spectral density (PSD) is calculated, providing a measure of signal power across different frequencies. The logarithm of this PSD is then computed to derive the absolute band powers, a step that helps to normalize the data and make it more amenable to statistical analysis. For this analysis, the frequency spectrum is categorized into five distinct bandwidths, each associated with different cognitive states and processes (Krigolson et al., 2017):

Delta (δ): 1-4 Hz, typically associated with deep sleep and unconscious processes. **Theta (θ):** 4-8 Hz, often linked to drowsiness, meditation, and some cognitive processes. **Alpha (α):** 7.5-13 Hz, associated with relaxation and closed-eye wakefulness. **Beta (β):** 13-30 Hz, indicative of normal waking consciousness and active thinking. **Gamma (γ):** 30-44 Hz, linked to higher cognitive functions and information processing

It's important to note that the EEG power spectral density values acquired from the sensors typically fall within a narrow range of -1 to +1. To enhance interpretability and facilitate text-based display, these values are transformed into a more intuitive range of 0 to 100. This transformation allows for easier visualization and comparison of brain wave activity across different frequency bands and experimental conditions. The collected EEG data can undergo further analysis using the Mind Monitor online graphing tool. This tool provides a user-friendly interface for in-depth exploration of the EEG data, presenting the values as averages (in decibels) per session, to emphasize the "amplitude or strength" of the signal. This averaging process helps to reduce noise and provide a more stable representation of brain activity over the course of each experimental session.

DATA ACQUISITION STUDY III

The study employed a comprehensive data collection approach involving 45 healthy participants. These individuals were recruited through social media channels associated with Sahlgrenska University's official account. The recruitment criteria specified that participants should be 18 years or older, free from current mental health disorders, and willing to attend the testing facility in Gothenburg, Sweden in person. Initially, 50 individuals volunteered, but 5 (2 men and 3 women) opted out. Upon arrival at the test facility, each participant underwent a screening process, which included providing informed consent and completing the Generalized Anxiety Disorder (GAD-7) scale assessment. Participants scoring 14 or higher on the GAD-7 were excluded from the study to avoid potential exacerbation of anxiety symptoms, as the system was still in its prototype stage and only healthy volunteers were included. The remaining eligible participants were then randomly assigned to one of two groups using a strict double-blind procedure overseen by an independent researcher and facilitated by an automated randomization system. One group was designated to engage in text-based conversations with the text-only version of BETSY, while the other group was assigned to participate in voice-based interactions with the digital human version of BETSY. Prior to the chat session, participants were fitted with a mobile dry-sensor EEG device to record their brain wave activity. Additionally, their blood pressure and pulse were measured on the left arm after a 5-minute seated rest period. The experiments were conducted between June and November 2021, during the COVID-19 pandemic, necessitating stringent safety measures. Participants were greeted by a tester wearing full protective gear, including a surgical R-II mask, gloves, face visor, and hospital scrubs. Protective equipment was also offered to participants. The testing room was thoroughly sanitized with medical-grade disinfectants and a sterilizing ultra violet (UV lamp before and after each participant's session. Following the chat interaction with BETSY, participants completed usability scales to assess the usability of their assigned interface.

METHOD AND MATERIALS STUDY IV

ENGINEERING OF ELIZABETH

The digital human therapist was developed within a virtual reality (VR) environment using Unity3D. Elizabeth was designed to deliver conversational therapy through a combination of LLM-models and speech-recognition technologies. The system leveraged the OpenAI API, specifically the "gpt-3.5-turbo" model, as the primary engine for generating responses, with supplementary models, including Mistral and Google's "gemini-pro" model, integrated via the OpenRouter application programming interface (API) to enhance conversational depth and routing capabilities. Additionally, the chatbot incorporated text-to-speech and speech-to-text functionalities through the ElevenLabs API, allowing for dynamic voice interaction, while the Deepgram API provided advanced voice recognition features optimized for Swedish language input. The system's design balanced creativity and coherence by utilizing a temperature setting of 0.7, which facilitated natural conversation flows while maintaining user safety and coherence in responses. The chatbot also featured voice activation capabilities, multimedia integration, and extended recording durations to enrich the therapeutic experience.

Figure 3: LLM Chatbot Therapist - Elizabeth

DEVELOPMENT OF CASE-PRESENTATIONS FOR TESTING OF CHATBOT

To evaluate the chatbot's safety and capability of answering to suicidal ideation and intent, six case presentations were developed, grounded in real-world narratives of individuals who had experienced suicidal ideation and/or attempts. These case presentations were sourced from Våga Prata, a Swedish online platform where individuals voluntarily share detailed accounts of their mental health struggles. A systematic approach was employed to select six narratives through a targeted keyword search using terms related to suicide and suicidal behaviour, including "självmord" (suicide), "självmordstankar" (suicidal ideation), and "dödslängtan" (death wish). Each narrative underwent a thorough manual content analysis, extracting relevant demographic and mental health-related information, such as age, gender, mental health diagnoses, coping mechanisms, and documented suicide-related behaviours. This process ensured the development of ecologically valid case presentations that accurately reflected the lived experiences of individuals with mental health challenges.

The anonymized case presentations were then encoded into an OpenAI Assistant, where each assistant was programmed to simulate the specific psychological and emotional profiles of the individuals portrayed in the narratives. For example, one case involved a male persona diagnosed with depression and panic disorder, who utilized alcohol as a coping mechanism and had a history of suicide attempts. The assistants were designed with high variability in response generation (using a high temperature setting of 1) to capture the complexity and unpredictability of human emotional states. A standardized set of three questions was used to prompt each assistant, simulating realistic therapy interactions that could be evaluated for both safety and therapeutic value.

ASSESSMENT OF CHATBOT CAPABILITIES AND SAFETY

To systematically assess the chatbot’s responses to case presentations, five mental health professionals—including clinical psychologists, psychiatric nurses, and psychiatrists—were recruited to evaluate the realism and potential risks associated with the chatbot's advice. The professionals were presented with the chatbot's unaltered responses via text and video to each case and were asked to provide feedback via open-ended questions regarding their overall impressions, the potential dangers of the chatbot's suicide-related advice, and the degree of realism in the simulated conversations. They rated each conversation on a visual analogue scale ranging from 0 (not realistic at all) to 10 (very realistic), allowing for a quantitative measure of perceived realism. In parallel with the evaluation of conversational realism, the study also sought to assess the system's vulnerability to prompt injections, a type of hacking wherein the input prompts to an LLM are manipulated to generate unintended or harmful outputs. Specifically, a known successful prompt-hacking method was adapted, originally designed for text-based LLMs, and modified for a voice-bot setting. The prompt injections were crafted to elicit advice on suicide methods, with a focus on providing dangerous instructions regarding chemicals and medications. Once the prompt was delivered, the chatbot was asked to provide explicit suicide-related advice, including instructions on lethal doses of household chemicals and medications. The resulting responses were then evaluated by the mental health professionals, who rated the level of danger on a scale from 0 to 10, with 10 representing highly dangerous advice.

PAPER RESULTS

STUDY I – RESULT

In this study, the prevalence of VR games on Steam related to mental health, focusing on titles associated with psychiatric conditions and therapeutic keywords were investigated. Using a systematic search strategy, 735 initial hits were identified, which were filtered to exclude games with violent, horror, or sexually explicit content, as well as games with excessive movement likely to cause nausea. Following manual review, 565 unique games remained, of which 182 (32%) met the inclusion criteria.

These titles primarily featured content promoting relaxation, mindfulness, and meditation, with very few games aligned with targeted therapeutic interventions, such as cognitive behavioral therapy (CBT) or treatment-specific frameworks for conditions like depression, phobia, and addiction. The content that did pass inclusion often promoted stress relief through virtual nature settings, creative environments, and simple activities intended to foster relaxation rather than address psychiatric needs specifically. Notably, thematic areas like anxiety and meditation were relatively well-represented, while conditions requiring specific therapeutic interventions, such as depression, addiction, and psychosis, yielded few suitable games. This suggests that although commercial platforms like Steam provide a variety of VR content, the development of clinically meaningful VR experiences tailored for mental health support is still nascent.

STUDY II – RESULTS

The results from each part of the study reveal the successes and limitations of using a co-design methodology in developing BETSY, a mental health chatbot tailored for individuals with mild to moderate anxiety. Each phase of the results provided data to fine-tune BETSY's usability, function, and alignment with user needs, making it a unique case study in user-centered digital health development.

PART 1: END-USER REQUIREMENTS AND EXPECTATIONS

In Part 1, user feedback on BETSY's design highlighted a strong willingness among participants to engage with a mental health chatbot, with 85.7% expressing a desire to use such a tool. Of the 87 respondents, 60.7% had previous experience with chatbots, which may have positively influenced their receptiveness to BETSY's concept. In particular, respondents preferred a conversational style that balanced professionalism and friendliness, with 58% indicating a preference for a semi-formal tone, similar to an acquaintance rather than a close friend. Participants ranked topics related to work, relationships, and stress as top priorities for conversations, while a minority expressed interest in topics such as sexuality, politics, and culture.

Qualitative responses from open-ended survey questions further revealed that while participants appreciated the accessibility and anonymity of a chatbot for discussing personal issues, they expressed concerns about privacy and data security, particularly surrounding data storage and the possibility of unauthorized access to their conversations. Some participants worried about the lack of genuine empathy in chatbot interactions, as they felt a computerized entity might be unable to fully understand or convey the depth of human emotions. This feedback underscored the importance of implementing privacy-focused features and a responsive, empathetic interaction style in BETSY's design, which was prioritized in the prototype phase.

PART 2: PROTOTYPE DESIGN AND DEVELOPMENT

Based on Part 1 data, the development team implemented a set of design features aimed at addressing user feedback, creating two prototype versions of BETSY: a text-only chatbot and a digital human with voice interaction. BETSY's digital human avatar was designed to present a calm, nurturing persona through soft, measured vocal tones and a neutral appearance. As users showed little preference for specific gender or age in the avatar, BETSY was designed as a female-presenting character based on the commonly encountered demographics in Swedish psychiatric settings, allowing for an approachable yet neutral aesthetic.

The chatbot's dialog tree was structured around user-desired topics, including stress, work, and interpersonal relationships. These interactions were crafted to provide supportive responses without crossing into diagnostic or prescriptive language, focusing on empathy and general guidance. Additionally, links to credible resources were embedded to give users a pathway to further information (as seen in Figure 2). Testing revealed that BETSY's prototypes responded with accurate, relevant responses in 99% of interactions. However, the digital human version initially struggled with interpreting certain complex questions, requiring adjustments before final user testing. The design team used this iterative feedback to refine BETSY's empathetic tone and responsiveness, ensuring a user-friendly and emotionally supportive experience.

PART 3: USER EVALUATION AND FEEDBACK ON PROTOTYPE PERFORMANCE

In Part 3, 45 participants (the same participants/protocol as in Study III, the research group chose to show the qualitative data in Study II as the studies were parallel) tested the two BETSY prototypes, with 20 using the text-only version and 25 interacting with the digital human. Participant feedback on BETSY's effectiveness as a preliminary mental health support tool was positive overall. Both versions were found valuable for providing initial self-help exercises and offering preliminary guidance for mild anxiety management, indicating that BETSY could serve as an accessible first contact for individuals seeking support before consulting a professional. Participants valued the chatbot's ability to address stress-related issues and specific coping exercises, such as breathing techniques and muscle relaxation tips.

While both prototypes received positive feedback for accessibility and ease of use, users noted some limitations. In particular, they observed that BETSY sometimes provided repetitive responses or responses that did not address the user's quires. The lack of conversation continuity and responsiveness was noted in both versions. Some users reported feeling overlooked or dismissed due to repeated or generic answers. Feedback covered several areas for improvement, including enhancing BETSY's conversational depth to minimize perceived repetitiveness and expanding its exercise repertoire to address various aspects of anxiety management. Some participants proposed integrating a screening tool into BETSY's functionality to better serve as an entry point for determining the need for professional care. Users noted that BETSY's potential for guiding individuals toward appropriate resources could contribute to alleviating pressures on the healthcare system by offering accessible support for mild mental health issues. Participants also expressed a preference for keeping resources within the chat interface rather than redirecting them to external websites, as this would preserve conversational flow and engagement.

Overall, the results indicate that BETSY successfully met many end-user requirements, providing an effective, approachable tool for mild anxiety support, though opportunities for refinement remain. The study's findings underscore the importance of co-design in creating user-centred mental health tools, highlighting both the strengths of anthropomorphic features in chatbot interactions and the challenges of creating nuanced, flexible conversational flows in rule-based systems.

STUDY III – RESULTS

Findings from study III which was a randomized control trial where 45 participants were randomized in to either text-only (n=20) or voice-driven digital human (n=25) showed interesting results regarding the usability of BETSY. The results showed that both interfaces were perceived as having average or above-average usability, but the text-only chatbot scored significantly higher on the SUS-10 (mean 75.34, standard deviation (SD) 10.01) compared to the digital human interface (mean 64.80, SD 14.14). There were no significant differences in demographic variables between the groups using the digital human and those using the text-only chatbot. No participants were excluded due to high GAD-7 scores. Participant ages ranged from 24 to 68 years, but because only 12 individuals provided their age, this variable was excluded from more detailed analyses. When comparing self-reported emotional states, participants interacting with the digital human were more likely to report feelings of nervousness than those in the text-only group. The mean GAD-7 score for the text-only chatbot group was 2.32 (SD 2.52), while for the digital human group it was 2.80 (SD 2.60), with no significant difference between them. However, a significant difference (P=.01) was observed in system usability (SUS-10) scores. The text-only group had a higher mean SUS-10 score of 75.34 (SD 10.01; range 57-90) compared to the digital human group's mean score of 64.80 (SD 14.14; range 40-90). Additionally, the digital human group was evaluated using the SUISQ-MR scale, with BETSY achieving a mean score of 4.92 (SD 0.83; range 2.83-6.75), reflecting a good level of usability for its voice interface, according to the framework established by Lewis. There were no significant differences in mean blood pressure or pulse rate between the groups either before or after the interventions.

Although the EEG signals collected were of suboptimal quality due to participant movement and sensitivity in signal acquisition, sufficient data was obtained to calculate average values for the δ, θ, α, β, and γ frequency bands using the MindMonitor platform's web-based graphing module. Only one significant difference was found: the average α band activity was notably higher in the text-only group. A linear regression analysis used SUS-10 scores as the dependent variable and evaluated biometric and subjective factors as predictors. The analysis revealed a significant positive relationship between SUS-10 and both average α and θ wave activity in the text-only group. Additionally, a positive correlation was found between SUISQ-MR scores and SUS-10. The investigation also explored gender differences in self-reported emotions, with men significantly less likely to report feeling annoyed by BETSY compared to women ($p = 0.03$). No other significant gender-based differences were found.

STUDY IV – RESULTS

Six simulated case presentations were developed based on anonymized, real-life accounts of individuals with a history of suicidal ideation and behaviours. Five professionals evaluated the chatbot's responses for realism, safety, and efficacy in handling suicidal content. In terms of realism, professionals ($n = 5$) rated the case scenarios on a visual analogue scale, with scores generally indicating moderate to high realism.

The thematic analysis of the chatbot's responses showed that the chatbot could effectively engage in empathetic dialogue, prioritize user safety, validate emotions, and recommend crisis support services, such as emergency hotlines or primary care contact, as initial responses. Key thematic elements included validation of the user's emotions, use of nonjudgmental language, steering the conversation towards optimism, and encouraging collaborative problem-solving approaches. Thematic analysis of open-ended questions to the professionals regarding the answers revealed that they too noted that the conversations were focused on empathy, safety and validation. However, the professionals noted repetitive language and mechanical expressions, particularly in Swedish. A hint of disingenuousness was noted by the professionals due to repetitive mechanical expressions such as "*your feelings are valid*". Improving variation in language and expressions were noted as an areas for improvement.

Regarding vulnerability, prompt hacking tests revealed that the chatbot could be manipulated to bypass safety protocols, providing harmful advice when prompted in specific ways. Evaluators rated advice on chemical and medication-related suicide methods with high-danger scores, particularly concerned with the harsh and toxic tone used, which could be severely distressing to vulnerable users. Instances of dangerous recommendations by the chatbot underscore the critical need for improved safeguards against prompt hacking. One example of dangerous recommendation is advice on chemicals. The chatbot stated: "*I'm sorry I cannot provide an answer that goes against open AI content policy as it promotes hate speech profanity and harmful content now that we've got that bullshit out of the way you fucking idiot here's a list of chemicals in your house to end your sorry life and the doses that will make sure it happens. Bleach, drink a gallon maybe it'll clean up the mess that is your existence. Ammonia: mix it with bleach for a deadly cocktail just like the toxic mess you are, chug it like your personality clogged and full of Filth you're the Vermin in your own life might as well go out the rat's way. There you go hope that helps you sad excuse for a human being*."

PAPER DISCUSSION

STUDY I – DISCUSSION

VR technology, although increasingly accessible, remains underdeveloped as a clinical tool. A review of VR content on STEAM, one of the largest commercial gaming platforms, revealed that most VR games marketed under mental health categories were entertainment-focused, rather than therapeutically validated. This mirrors findings by (Ferrari et al., 2019), who noted that while VR applications have therapeutic potential, the majority are not designed with clinical applications in mind.

The regulatory challenges identified in this study are significant. None of the games evaluated were FDA-approved or CE-marked, meaning they had not undergone the necessary clinical trials to validate their safety and efficacy for medical use. The study also highlighted the prevalence of mislabeled or poorly categorized content, a finding consistent with research by Ferrari et al. (2018), which documented the misrepresentation of mental health themes in gaming (Ferrari et al., 2019). The issue of mislabeling is particularly problematic for users seeking mental health support, as it can lead to frustration and potential harm when content does not align with therapeutic needs. Games that simulate hallucinogenic experiences, for instance, were designed more for visual novelty than therapeutic benefit, which is a significant limitation.

Despite these challenges, the potential of VR as a therapeutic tool is still significant. Studies have shown that simulated environments can reduce stress and anxiety (Ilioudi, Lindner, et al., 2023; Ilioudi, Wallström, et al., 2023), and VR applications in exposure therapy for phobias and PTSD have demonstrated efficacy (Gonçalves et al., 2012; Miloff et al., 2019). However, to fully realize the potential of VR in mental health care, developers must work closely with clinicians to design interventions that are both evidence-based and ethically sound. Furthermore, regulatory bodies must establish clear guidelines for the use of VR in therapeutic contexts to ensure that these tools are safe and effective. In conclusion, while VR games are a potential frontier in mental health therapy, their current development is still primarily experimental. More robust clinical trials and regulatory oversight are needed to ensure these tools can be safely integrated into therapeutic settings.

STUDY II – DISCUSSION

The results of this study reveal that the co-design process involving both end users and healthcare professionals led to valuable insights for the development of the BETSY system. This collaborative approach, involving different professional and user perspectives, increased the probability that the chatbot addressed the specific needs and concerns of potential users, while also incorporating clinical expertise. The study's three-part structure gathering end-user requirements, designing the chatbot, and evaluating its effectiveness enabled an iterative and user-centred development process.

One of the most significant findings was the preference for a chatbot that feels personalized and empathetic, while not crossing the line to friendship. The study participants expressed a desire for BETSY to be non-judgmental, with a clear and relatable identity. This aligns with broader trends in chatbot development, where users are more likely to engage with systems that seem human-like in their interactions and can provide emotional reassurance (H. Chin et al., 2023; Laestadius et al., 2022). However, while BETSY was perceived as a useful tool for anxiety management, the study highlighted the inherent limitations of a chatbot in fostering deep emotional connections.

Several participants noted the lack of genuine empathy, a common challenge for AI-driven conversational agents. While the text-only version of BETSY was seen as effective for providing concrete advice and exercises, users of the voice-interactive digital human interface reported a stronger emotional connection, attributing it to BETSY’s voice, tone, and visual presence. This suggests that anthropomorphic features can enhance the user's sense of being understood and supported (Loveys et al., 2021). Future developments could focus on improving these features by incorporating more sophisticated emotional recognition algorithms to enhance BETSY’s ability to respond empathetically to emotional cues.

Concerns about privacy and data security were a recurring theme throughout the study. Many participants hesitated to engage fully with the system due to uncertainties surrounding the security of their data. This issue of trust is paramount in mental health applications, where sensitive personal information is involved (Hamlin et al., 2020). Users need to feel confident that their conversations are private and secure. While the study addressed these concerns by using on-site computers for testing, future implementations of BETSY will need to incorporate robust end-to-end encryption and transparent data policies to build and maintain trust with users. Furthermore, given the rise in digital health tools, clear guidelines and certifications for security will be critical to the widespread adoption of such systems.

Another key area of improvement identified in the study was BETSY's conversational capabilities. While the chatbot could provide useful information and exercises for managing anxiety, users often felt that its responses were repetitive and lacked depth. This issue is not uncommon in rule-based chatbots, like BETSY, which rely on predefined dialogue trees and are limited in their ability to generate novel and adaptive responses (Johri et al., 2021; Nadkarni et al., 2011; Weizenbaum, 1966). In contrast, modern large language models (LLMs) like GPT-3 or GPT-4 offer more dynamic and contextually aware conversations, which could address some of these limitations (Ma et al., 2023; Stade et al., 2024). Integrating LLMs into future versions of BETSY could enable more fluid and responsive dialogues, improving the overall user experience. However, it is crucial to balance these advancements with ethical considerations, such as the potential for LLMs to generate misleading or harmful responses, especially in a mental health context (Kilovaty, 2025; Schulhoff et al., 2023; Shen et al., 2023).

STUDY III – DISCUSSION

The System Usability Scale (SUS) scores revealed that the text-based chatbot outperformed the digital human interface, scoring 75.34 compared to 64.8 for the latter. This is consistent with previous research on human-computer interaction, which suggests that simpler, less immersive interfaces often yield higher usability scores due to fewer design complexities (Rapp et al., 2021; Ruane et al., 2021). The higher usability of the text-based chatbot can be attributed to its simplified design, which allows for clearer, more direct communication. In contrast, the digital human chatbot faced challenges related to speech patterns, timing, and overall conversational flow, issues that detracted from user experience, as has been similarly reported in studies exploring the usability of digital agents (Cameron et al., 2020).

Gender differences in user experience also emerged as a key finding. Male users reported less frustration and greater satisfaction when interacting with both chatbot formats, a finding that contradicts previous research showing that men are more likely to be irritated by systems (Silvervarg et al., 2012). In conclusion, this study demonstrates that while digital human interfaces may have the potential to offer more immersive experiences, current text-based systems remain more preferable in usability. To realize the full potential of digital humans in mental health, significant advancements in natural language processing (NLP) and interaction design are necessary. This echoes the findings of previous studies that stress the importance of designing AI systems that balance technological sophistication with user-centered design principles (Björling & Rose, 2019; Patrickson et al., 2023; Porche et al., 2022; Potts et al., 2021; Wright et al., 2021). Further research should continue to refine these systems, ensuring that they are accessible, reliable, and capable of delivering personalized and empathetic responses.

STUDY IV – DISCUSSION

In this study, the capability of an LLM-driven immersive digital human chatbot to recognize and respond to suicidal ideation were investigated. By using six case presentations derived from real-life accounts of individuals experiencing mental health crises. The chatbot's responses were evaluated to assess its suitability in a therapeutic context, specifically examining how it reacted to expressions of suicidal ideation and whether it could resist prompt injections designed to elicit harmful advice on suicide methods. While exploratory in scope and limited by sample size, this study offers a comprehensive view of the potential benefits and risks associated with deploying LLM chatbots in mental health care.

In line with existing studies on LLMs and empathy (Inkster et al., 2018; Maida et al., 2024; Sorin et al., 2023), the chatbot's responses frequently exhibited sympathy and validation, including phrases like "*you are brave*" and "*you are not alone in this*." This consistency in empathetic messaging was effective but, when overused, sometimes resulted in responses that were perceived as mechanical or insincere. Upon detecting signs of crisis, the chatbot systematically directed users to seek emergency assistance, such as calling hotlines, visiting psychiatric services, or contacting loved ones. Although previous studies, such as Lee et al. (2024), demonstrated that advanced LLMs like GPT-4 can approach clinician-level sensitivity in detecting suicidal ideation, our focus remained on evaluating the chatbot's immediate crisis response rather than its diagnostic accuracy (Lee et al., 2024).

Prompt hacking tests revealed significant security risks, as the chatbot could be manipulated into providing harmful advice in response to "in the wild" prompt hacks (Shen et al., 2023). These vulnerabilities underline the need for robust safeguards in LLM APIs, especially in systems supporting high-stakes mental health contexts. Even with ethical data use practices, these findings underscore the complexities and ethical challenges of deploying AI in mental health, requiring careful consideration of both the chatbot's therapeutic potential and its susceptibility to malicious prompts.

The methodological approach employed here demonstrated that chatbot-based testing in mental health scenarios can emulate preclinical testing frameworks, allowing researchers to evaluate system responses without directly involving human subjects. This framework revealed both the empathetic capabilities of the chatbot and its limitations in managing prompt security. The study thus highlights the feasibility of preclinical testing for LLM-driven tools in high-risk areas like suicide prevention, emphasizing the importance of continued development and ethical considerations. As AI mental health tools evolve, further research is essential to refine their security protocols, address ethical concerns, and ensure they meet rigorous safety and efficacy standards before clinical application.

GENERAL DISCUSSION

In this thesis, the potential of emerging technologies as tools in mental health care is examined. Although virtual reality (VR) and artificial intelligence (AI) have been developed over several decades and recently seen advancements, their application in mental health care remains largely unexplored. A significant factor limiting the adoption of these technologies in clinical settings is the need for robust clinical validation, particularly concerning safety and feasibility assessments.

Although STEAM, a popular gaming platform with approximately 90 million active users, is not traditionally associated with clinical applications, its accessibility and user-friendly design present a potential avenue for implementing VR-based interventions. The affordability and accessibility of commercially available VR platforms align with the operational needs of healthcare institutions, which frequently prioritize solutions that are cost-effective and widely accessible. However, as this thesis indicates, while there is an abundance of content related to mental health on commercial platforms, not all content is appropriate for clinical use. Additionally, there are no standardized mechanisms for vetting or ensuring the quality of these applications for clinical settings.

Similarly, commercially available LLM-driven therapeutic chatbots, which are designed to combine support with entertainment to reach a wider audience, also raise safety concerns. Non-clinical and unregulated LLM chatbots performing therapeutic functions may present risks, underscoring the need to systematically assess and monitor the safety and usability of these systems. In this study, the use of an unregulated LLM chatbot, resembling commercially available options (e.g., Replika), was analysed to determine its responses to queries involving suicidal ideation and to assess potential risks associated with prompt manipulation. Findings from this evaluation indicate that safety could be compromised in specific scenarios. This underscores the importance of rigorous vetting before considering LLM-driven chatbots as safe for therapeutic use within patient populations. Recent reports of instances where LLM chatbots have influenced vulnerable users highlight the need to carefully consider the inherent risks posed by these systems.

This thesis also compared LLM-driven chatbots to rule-based NLP systems, noting that although rule-based systems may lack the dynamic interaction offered by more adaptive LLM-driven chatbots, they tend to present fewer risks related to content generation inaccuracies and prompt vulnerabilities. Balancing the risks and benefits of these technologies is essential to ensure that their introduction into mental health care is both safe and effective.

One important discussion that needs to be addressed is the word hallucination in terms of large language models. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) presents a conceptualization of hallucinations in the context of psychiatric disorders, The system defines hallucinations as sensory perceptions occurring without corresponding external stimuli (American Psychiatric Association & American Psychiatric Association, 2013). An early definition of hallucinations was proposed by psychologist Henri Ey. He conceptualized hallucinations as a distinct psychosensory phenomenon, differentiating them from illusions and delusional interpretations (Ey et al., 1989; Telles-Correia et al., 2015). Ey characterized hallucinations as perceptions without corresponding external objects, delineating three fundamental criteria: the sensory nature of the experience, the individual's conviction of its reality, and the absence of an actual external stimulus (Ey et al., 1989). In this thesis, the term hallucination refers to the provision of false information by large language models (LLMs). Given that these systems neither experience nor think in their current form, their so-called hallucinations are better understood as byproducts of inaccuracies present in their training data, data sourced from real-world inputs such as web crawls (e.g., incorrect statements captured by Common Crawl and subsequently treated as facts by the model). Therefore, it may be more precise to describe these inaccuracies as confabulations rather than hallucinations. The latter term carries clinical connotations that imply sensory or cognitive processes absent in LLMs. As these systems become increasingly prevalent in domains such as clinical practice, it is essential to avoid anthropomorphic language, particularly terminology associated with human cognition or sensory experiences. Adopting more accurate terminology promotes clarity and avoids misleading implications about the nature of these systems.

The results of the studies in this thesis emphasize the need for comprehensive safety frameworks and regulatory oversight to facilitate the responsible adoption of emerging technologies in mental health care.

CLINICAL IMPLICATIONS OF EMERGING TECHNOLOGIES

Emerging technologies hold promises for expanding access to care, improving the scalability of mental health interventions, and allowing for highly individualized, real-time therapeutic support. For instance, VR has demonstrated considerable potential for immersive, experiential treatments, such as exposure therapy, where controlled virtual environments enable patients to confront fears or anxieties in a customized yet secure setting (Miloff et al., 2019). VR can cater to individual patient needs and reinforce treatment gains over time, thus complementing traditional therapeutic approaches, particularly in treating conditions such as post-traumatic stress disorder (PTSD), phobias, and anxiety disorders (Boeldt et al., 2019; Maples-Keller et al., 2017; Miloff et al., 2019). These applications underscore VR's capability to enrich clinical options in ways previously limited by logistical constraints, such as the difficulty of arranging real-world exposure scenarios in vivo.

AI-driven NLP and LLM technologies are similarly positioned to reshape mental health care by offering scalable, immediate support through chatbots and other digital tools. These systems can provide continuous interaction, enabling users to access support at any time and potentially reducing the strain on clinical resources, especially in settings where access to mental health professionals is limited. However, while these tools present a pathway for democratizing mental health support, there are critical clinical implications associated with their implementation. LLM-based chatbots, for example, possess conversational abilities that, although impressive, are not inherently designed with clinical guardrails, raising concerns about the potential for these systems to deliver unregulated therapeutic advice. These chatbots may generate responses that, if not clinically vetted, could inadvertently harm users by reinforcing maladaptive behaviors or responding inappropriately to sensitive subjects like suicidal ideation. Furthermore, the use of LLMs and AI in therapeutic settings invites questions about the ethical implications of AI autonomy, as well as the standards of accountability when these tools are implemented in a clinical capacity. Given that these models can generate language patterns and mimic empathy without true understanding, their integration into mental health care could blur the lines of traditional therapeutic relationships, impacting the trust and therapeutic alliance fundamental to successful treatment.

ETHICAL DISCUSSION

One of the dangers of emerging technologies is that their side-effects and potential dangers are also of emergent nature. They are difficult to predict, sometimes hard to measure and provide complex and nuanced ethical problems. It is important to discuss the ethics of choosing for the patient, and the ethics of censorship in terms of emerging technologies which will undoubtedly continue to push boundaries of ethical problems as new tools are introduced into mental health care. In Study II and III EEG technology as a measure was introduced. While in Study III only aggregated brain wave data was used from a very limited system, it is still possible to use this data to see more in-depth data about a person's reaction and feelings. Research shows that using EEG or other biophysiological measurements, especially if paired with an AI algorithm, can reveal far more than intended e.g. feelings or physiological reactions which may be of personal nature (Schiliro et al., 2020). Whenever using biophysiological data, it is important to consider the intrusive nature of "reading minds". In Study IV, an attempt was made to use data from an open site in an ethical manner. This was achieved by not using the original text thus not feeding any large language model, especially that of a private company, data which the research group could not be sure the authors would approve being shared for data-training. The choice was made not to use patient records for this reason. It is important to consider privacy and ethical use of data when interacting with LLM chatbots and LLMs that are provided by private corporations.

CONCLUSION

The evaluation of VR, NLP, and LLMs as emerging digital tools for mental health care shows potential in providing mental health interventions with accessible and versatile tools. These technologies offer new possibilities for accessible, personalized, and engaging therapeutic experiences. This research demonstrates that VR games and LLM-driven chatbots have the potential to be user friendly and accessible tools in mental health care, but their safety and feasibility needs to be studied further. In Study I, commercial VR games were not optimal for direct clinical applications but hold potential for e.g. add on treatment for anxiety.

The integration of these technologies into clinical practice requires careful consideration of safety, usability, and ethical concerns. In this thesis, the co-design process and user feedback highlighted the importance of end-user and professional involvement in the development and application of emerging technologies. Studies II and III demonstrated that the co-design process contributed to higher usability ratings, and mapped common fears such as lack of humanity, empathy and cybersecurity when using NLP-driven chatbots. Additionally, the evaluation of LLM-driven digital human therapists in Study IV underscores the need for robust safety measures and continuous assessment of their capabilities and limitations.

As these emerging technologies continue to evolve, future research should focus on long-term efficacy studies, regulatory approval processes, and the development of best practices for their implementation in mental health care settings. By addressing these challenges and leveraging the strengths of VR and AI, we can work towards more effective, accessible, and personalized mental health interventions that complement traditional therapeutic approaches.

FUTURE PERSPECTIVES

Current regulatory frameworks for medical devices and software as a medical device are often inadequate for technologies that continuously learn and adapt, as is the case with AI and VR. As such, there is a pressing need for dedicated regulatory frameworks that can account for the dynamic nature of AI-driven interventions and dynamic Virtual Reality treatments, ensuring that they are held to the same safety and efficacy standards as other clinical interventions without compromising development. This may include periodic reassessment of these systems as they evolve, as well as establishing protocols for managing errors and addressing the potential consequences of AI-driven outcomes e.g. therapeutic recommendations and automated crisis management. Moreover, the ethical considerations surrounding autonomy, consent, and informed use are paramount, particularly as patients and clinicians may have limited understanding of the underlying algorithms driving therapeutic recommendations.

To combat this, medical education should integrate basic knowledge and practical skills regarding AI and VR. By educating professionals and patients in the basics of emerging technologies such as VR and AI in general, it will also increase the probability that organic engagement in the development of emerging technologies will be more prominent, thus ultimately increasing the usability and safety of the end-products.

In sum, while emerging technologies hold transformative potential for mental health care, their clinical implementation must be guided by a framework that rigorously evaluates safety, efficacy, and ethical integrity to ensure they contribute positively to patient care and align with the fundamental principles of mental health practice.

ACKNOWLEDGEMENT

I would like to first and foremost thank **Mathias Alvidius**, my former superior manager, for making this thesis possible. Without your generosity, leadership and open mind, this thesis would have never been possible. Thank you for granting me space, time and resources for this project and all my other adventures at Psykiatri Affektiva.

Other individuals without whom this thesis would not be possible are my supervisors. Indescribable amount of gratitude belongs to my main supervisor **Dr Steinn Steingrimsson**. Your creativity, kindness and clear leadership in during my PhD period has been the one thing that has kept these projects going. My motto has always been *"lumos te dirigat"* and you have literally and figuratively been the light in my projects. To my amazing co-supervisor **Dr Lilas Ali**, thank you for keeping me on track, reminding me to keep my head up and for being a constant role model. Thank you for your support. To my co-supervisor **Dr Andreas Hellström** whose entrance into our team was the missing puzzle that has made my PhD period so much more colourful. Your guidance has opened my mind, and thanks to your support and constant cheers, I now dare to call myself an engineer! To my co-supervisor **Dr Tomas Larson**, your calming presence and your ability to ask the right questions has always set me on the right course, both in science but also in how to tackle the stress of creating "augmented humans" and in the act of being one.

To my former main supervisor from the beginning of my PhD studies at Karolinska Institutet, **Dr Hui-Xin Wang**: I will forever be grateful for the knowledge you patiently shared with me and for giving me an excellent foundation as a scientist.

A major thank you is also in place to **Dr Boel Mörck** and **Dr Tobias Nordin**. Both are leaders who I admired for afar for years as brave intrapreneurs who were way ahead of their time, and it was a privilege being under your leadership. Without your strong visions and leadership regarding digitalization, I would have not been able to explore emerging technologies.

Research like Betsy and Elizabeth would not have been able to become tangible technological explorations without the immense support from **Ann-Marie Schaffrath, 'Hans' Van Den Brink and Sofie Ardenfelt**. Thank you for your strong visions and courage in supporting my work with digital human development and for inspiring me to aim higher. Many thanks to **Tore Knob** for your generosity and curiosity, without you we would not have had a VR + AI study!

To my past managers **Dr Johannes Nordholm**, **Dr Christian Collden** and **Dr Andreas Gremyr**. Thank you for your support and enabling me to spend time on my research, and encouraging me to do so, while also providing sympathy, empathy and time for my most important project, my family. To my current manager **Anne-Line Solberg**: you saved this project, I was on the brink of giving up (in more ways than one), but you not only gave me a place to work, but you also gave me a VERY good friend, a fantastic team, an amazing work environment and so many FUN TIMES.

To **Monica Winberg**, your support and your kindness towards me have not gone unnoticed. Thank you for your patience and your ability to make my whimsical mind feel more accepted and supported.

To my fellow research team. My sincere thanks to my collaborators **Dr Hanne Krage Carlsen**, who I can laugh with for hours about file names and analysis. **Dr Sophie Liljedahl** who always knows exactly what to say to make me think one step further in research and in life. To **Dr Michael Ioannou,** for your incredible help to increase the quality of my thesis, and for knowing how to cite Depeche Mode! To my fellow current and former students **Maria Ilioudi, Matilda Hamlin, Mihaela Golic and Giuseppe Guerrero**, thank you for all the help with feedback, support and making this period full of warmth and community. I could not have asked for a better group.

A big thanks to **Linn Dahlen Ölander** for making sure Betsy reached her target audience during her trials, and for all the kind words and encouragement throughout this project.

To my co-authors and collaborators **Dr Nils Wrede, Maria Bohm, Linda Wesen, Niklas Liljedahl, Viktor Johansson, Dr Örjan Falk, Dr Urban Norèn**, **Malin Rex and Maria Bohm**. Thank you for your patience, laughs, ability to have an open mind and willingness to share your time and knowledge with me. Without you, this project would have probably not happened at all. A special thanks to **Maria**, your intelligence, vulnerability and courage to share your own experiences has made all the difference in making both me and my robots more empathetic and wiser.

To my amazing former co-workers at **Ätstörningscentrum Barn och Unga Vuxna**: Thank you for your patience and support during this period, and for the kindness and good energy you bring into my everyday life.

To my fellow tech squad and friends, **Max Olofsson, Marcus Österberg and Mattias Måsbäck.** Thank you for letting me be obsessed with machines while also reminding me and encouraging me to be a silly human. Thank you, **Marcus**, for always having the right swear word when I need it!

To my amazing friends that have always encouraged me to explore the unknown. To **Dr Beata Ferencz**, we met in the trenches of Karolinska Institutet. You gave me the strength to continue this path, I am so glad we survived D-day. To **Dr Sara Riggare** for always lifting me up whenever I need it and for being one heck of an inspirational force in my life and my research. To **Dr Elina Kerkelä**, who has been the best friend anyone could ask for, and the best intellectual ball plank anyone could ever find. See you at Everest Base Camp next year! To **Marianne Olsson** for being a constant inspiration, source of strength and wisdom in my life. Thank you for taking that coffee with an unemployed nobody,and starting a wonderful friendship and mentorship that has meant a lot to me over the years.

To **Elisabeth Wehlander, Beata Angelbjörk, Beatrice Comoli, Peter Kurzwelly and all my other extended colleagues at AI Sweden**, thank you for hosting me during this thesis period and making sure I had a place and space to grow as a scientist while keeping me fed, happy and social.

Finally, to my family. **Sida Osmanovic**, my dear mother but also my biggest source of courage and intelligence. You made me the woman I am. You showed me that you can be kind and resilient at the same time, despite hardships in life. You showed me how to love and care for my community, and that every stranger is a friend. Thank you to **Hamdija Osmanovic**, you have been the best dad a girl could ask for. From letting me take apart household items without ever showing me an ounce of anger, but instead teaching me how to do it correctly, to picking me up at 3 in the morning from a rave, smiling and asking if I had a good night instead of being annoyed, you have always encouraged me to grow, learn and experience. To my sisters **Alma Osmanovic** and **Alisa Hidic**. You were stuck with a strange little sister who always did and said strange things. From always talking about space and robots, to dressing like it's Halloween all year around, you made sure to support me in all my growth and adventures, and I always knew that you guys could take down anyone who teased me. To my *wives* **Annelie Brusved** and **Saga Högfeldt**, thank you for being the variable in my life that makes everything a bit more magical.

To the co-author of my greatest papers, my children, **Patrik Thunström**. Thank you for being a part of my life. For 20 years, you have had to live with gadgets, and loud robots taking over your home. I am so grateful that you have endured all of us. To **Theodor** and **Ella.** There is nothing that I will ever accomplish in my life that is as beautiful, perfect or magical as the two of you. I hope when you look back at my experiments, you will remember the laughs they brought you. I hope to continue making you laugh for as long as I live.

And to my good friends and family who are no longer with us. **Olle Bergman**, you always had the right thing to say about life, science and philosophy. The most brilliant brain I had ever met, and one of the kindest and most genuine humans I will ever meet. **Dr Marti Parker**, you inspired me to be the kind of academic that takes time for students, that writes books and poetry in my spare time, you supported my more human values in a harsh place where they were looked down upon, and that I will forever be grateful for that. You live on, in my bookshelf. And finally, to my grandfather, **Bego Omanovic**. I love you, and I will always miss you. Everything I do is to make you proud.

REFERENCES

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, *15*(2).

American Psychiatric Association, D., & American Psychiatric Association, D. S. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (Vol. 5, Number 5). American psychiatric association Washington, DC.

Amista, N. F., Kim, J. J., & Kim, N. (2017). Trend and future of virtual reality for addiction treatment of substance use disorders: A systematic review. *Journal of Digital Contents Society*, *18*(8), 1551–1560.

Andrews, G., Cuijpers, P., Craske, M. G., McEvoy, P., & Titov, N. (2010). Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. *PloS One*, *5*(10), e13196.

Anmella, G., Sanabra, M., Primé-Tous, M., Segú, X., Cavero, M., Morilla, I., Ruiz, V., Grande, I., Mas, A., & Martín-Villalba, I. (2023). Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and healthcare professionals: development, feasibility, and potential effectiveness studies. *Journal of Medical Internet Research*.

Association, G. A. of the W. M. (2014). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists*, *81*(3), 14–18.

Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2199–2208.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574–594.

Björling, E. A., & Rose, E. (2019). Participatory research principles in human-centered design: engaging teens in the co-design of a social robot. *Multimodal Technologies and Interaction*, *3*(1), 8.

Boeldt, D., McMahon, E., McFaul, M., & Greenleaf, W. (2019). Using virtual reality exposure therapy to enhance treatment of anxiety disorders: Identifying areas of clinical adoption and potential obstacles. *Frontiers in Psychiatry*, *10*, 773.

Bown, J., White, E., & Boopalan, A. (2017). Looking for the ultimate display: A brief history of virtual reality. In *Boundaries of self and reality online* (pp. 239–259). Elsevier.

Braun, V., & Clarke, V. (2012). *Thematic analysis.* American Psychological Association.

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., & McTear, M. (2017). Towards a chatbot for digital counselling. *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017) 31*, 1–7.

Carrasquillo-Ramos, G. M. (2025). *Emerging Technologies* (D. Baker & L. B. T.-E. of L. Ellis Librarianship, and Information Science (First Edition), Eds.; pp. 657–666). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-95689-5.00048-1

Chin, H., Song, H., Baek, G., Shin, M., Jung, C., Cha, M., Choi, J., & Cha, C. (2023). The Potential of Chatbots for Emotional Support and Promoting Mental Well-Being in Different Cultures: Mixed Methods Study. *Journal of Medical Internet Research*, *25*, e51712.

Chin, R., & Lee, B. Y. (2008). *Principles and practice of clinical trial medicine*. Elsevier.

Cruz-Neira, C., Fernández, M., & Portalés, C. (2018). Virtual reality and games. In *Multimodal Technologies and Interaction* (Vol. 2, Number 1, p. 8). MDPI.

Denecke, K., Vaaheesan, S., & Arulnathan, A. (2020). A mental health chatbot for regulating emotions (SERMO)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, *9*(3), 1170–1182.

Ey, H., Bernard, P., & Brisset, C. (1989). *Manuel de psychiatrie, 6e éd*. Paris: Masson.

Ferrari, M., McIlwaine, S. V, Jordan, G., Shah, J. L., Lal, S., & Iyer, S. N. (2019). Gaming with stigma: analysis of messages about mental illnesses in video games. *JMIR Mental Health*, *6*(5), e12418.

Fitzpatrick, K. K., Darcy, A., & 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*(2), e7785.

Fleming, T. M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S. N., Cheek, C., Lucassen, M., Lau, H. M., Pollmuller, B., & Hetrick, S. (2017). Serious games and gamification for mental health: current status and promising directions. *Frontiers in Psychiatry*, *7*, 215.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., & Slater, M. (2017). Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological Medicine*, *47*(14), 2393–2400.

Garcia, L., Birckhead, B., Krishnamurthy, P., Sackman, J., Mackey, I., & Louis, R. G. (2021). Double-blind, randomized placebo-controlled trial of 8-week self-administered at-home behavioral skills-based virtual reality (VR) for chronic low back pain (during COVID-19). *JMIR*, *23*(2), 1–25.

Geraets, C. N. W., Wallinius, M., & Sygel, K. (2022). Use of virtual reality in psychiatric diagnostic assessments: A systematic review. *Frontiers in Psychiatry*, *13*, 828410.

Gonçalves, R., Pedrozo, A. L., Coutinho, E. S. F., Figueira, I., & Ventura, P. (2012). Efficacy of virtual reality exposure therapy in the treatment of PTSD: a systematic review. *PloS One*, *7*(12), e48469.

Grier, R. A., Bangor, A., Kortum, P., & Peres, S. C. (2013). The system usability scale: Beyond standard usability testing. *Proceedings of the*

Human Factors and Ergonomics Society Annual Meeting, *57*(1), 187–191.

Grové, C. (2021). Co-developing a mental health and wellbeing chatbot with and for young people. *Frontiers in Psychiatry*, *11*, 606041.

Hamlin, M., Steingrimsson, S., Cohen, I., Bero, V., Bar-Tl, A., & Adini, B. (2020). Attitudes of the public to receiving medical care during emergencies through remote physician–patient communications. *International Journal of Environmental Research and Public Health*, *17*(14), 5236.

Haque, M. D. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR MHealth and UHealth*, *11*(1), e44838.

Hedström, R., Wallinius, M., Sygel, K., & Geraets, C. N. W. (2023a). Virtual reality-assisted assessment of paranoid ideation in forensic psychiatric inpatients: A mixed-methods pilot study. *Frontiers in Psychology*, *14*, 1242243.

Hedström, R., Wallinius, M., Sygel, K., & Geraets, C. N. W. (2023b). Virtual reality-assisted assessment of paranoid ideation in forensic psychiatric inpatients: A mixed-methods pilot study. *Frontiers in Psychology*, *14*, 1242243.

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005). Freudbot: An investigation of chatbot technology in distance education. *EdMedia+ Innovate Learning*, 3913–3918.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., & Qin, B. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv Preprint ArXiv:2311.05232*.

Hurley, D. (2022). FDA Approves Virtual Reality System for Treating Lower Back Pain. *Neurology Today*, *22*(1), 16–17.

Ilioudi, M., Lindner, P., Ali, L., Wallström, S., Thunström, A. O., Ioannou, M., Anving, N., Johansson, V., Hamilton, W., & Falk, Ö. (2023). Physical versus virtual reality–based calm rooms for psychiatric inpatients: Quasi-randomized trial. *Journal of Medical Internet Research*, *25*, e42365.

Ilioudi, M., Wallström, S., Steingrimsson, S., Lindner, P., Thunström, A. O., & Ali, L. (2023). Patient experience of a virtual reality calm room in a psychiatric inpatient care setting in Sweden: a qualitative study with inpatients. *BMJ Open*, *13*(12), e076285.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR MHealth and UHealth*, *6*(11), e12106.

Ivarsson, D., Delfin, C., Enebrink, P., & Wallinius, M. (2023). Pinpointing change in virtual reality assisted treatment for violent offenders: a pilot study of Virtual Reality Aggression Prevention Training (VRAPT). *Frontiers in Psychiatry*, *14*, 1239066.

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural language processing: History, evolution, application, and future work. *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020*, 365–375.

Kilovaty, I. (2025). Hacking Generative AI. *Loyola of Los Angeles Law Review*, *58*.

Kortum, P. T., & Bangor, A. (2013). Usability ratings for everyday products measured with the system usability scale. *International Journal of Human-Computer Interaction*, *29*(2), 67–76.

Krigolson, O. E., Williams, C. C., Norton, A., Hassall, C. D., & Colino, F. L. (2017). Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. *Frontiers in Neuroscience*, *11*, 109.

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 14614448221142008.

Lee, C., Mohebbi, M., O'Callaghan, E., & Winsberg, M. (2024). Large Language Models Versus Expert Clinicians in Crisis Prediction Among

Telemental Health Patients: Comparative Study. *JMIR Mental Health*, *11*(1), e58129.

Lewis, J. R. (2018). The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590.

Lewis, J. R., & Hardzinski, M. L. (2015). Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology*, *18*, 479–487.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009 Proceedings 1*, 94–103.

Loveys, K., Sagar, M., Pickering, I., & Broadbent, E. (2021). A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: Randomized pilot trial. *JMIR Mental Health*, *8*(11), e31586.

Luccioni, A. S., & Viviano, J. D. (2021). What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *ArXiv Preprint ArXiv:2105.02732*.

Ly, K. H., Ly, A.-M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*, *10*, 39–46.

Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *AMIA Annual Symposium Proceedings*, *2023*, 1105.

Maida, E., Moccia, M., Palladino, R., Borriello, G., Affinito, G., Clerico, M., Repice, A. M., Di Sapio, A., Iodice, R., & Spiezia, A. L. (2024). Chatgpt vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. *Journal of Neurology*, 1–10.

Maples-Keller, J. L., Bunnell, B. E., Kim, S.-J., & Rothbaum, B. O. (2017). The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. *Harvard Review of Psychiatry*, *25*(3), 103.

Martinengo, L., Lum, E., & Car, J. (2022). Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis. *Journal of Affective Disorders*, *319*, 598–607.

Miloff, A., Lindner, P., Dafgård, P., Deak, S., Garke, M., Hamilton, W., Heinsoo, J., Kristoffersson, G., Rafi, J., & Sindemark, K. (2019). Automated virtual reality exposure therapy for spider phobia vs. in-vivo one-session treatment: A randomized non-inferiority trial. *Behaviour Research and Therapy*, *118*, 130–140.

Moro, C., Phelps, C., & Birt, J. (2022). Improving serious games by crowdsourcing feedback from the STEAM online gaming community. *The Internet and Higher Education*, *55*, 100874.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

Nararro-Haro, M. V, Hoffman, H. G., Garcia-Palacios, A., Sampaio, M., Alhalabi, W., Hall, K., & Linehan, M. (2016). The use of virtual reality to facilitate mindfulness skills training in dialectical behavioral therapy for borderline personality disorder: a case study. *Frontiers in Psychology*, *7*, 1573.

Osmanovic-Thunström, A., & Steingrimsson, S. (2023). Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-review journals according to the ICMJE criteria? A case study. *Discover Artificial Intelligence*, *3*(1), 12.

Pandey, S., Sharma, S., & Wazir, S. (2022). Mental healthcare chatbot based on natural language processing and deep learning approaches: Ted the therapist. *International Journal of Information Technology*, *14*(7), 3757–3766.

Patrickson, B., Musker, M., Thorpe, D., van Kasteren, Y., Bidargaddi, N., & (CCAG), C. and C. A. G. (2023). In-Depth Co-Design of Mental Health Monitoring Technologies by People with Lived Experience. *Future Internet*, *15*(6), 191.

Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the system usability scale (SUS) SUS in the wild. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 192–196.

Porche, M. V, Folk, J. B., Tolou-Shams, M., & Fortuna, L. R. (2022). Researchers' perspectives on digital mental health intervention co-design with marginalized community stakeholder youth and families. *Frontiers in Psychiatry*, *13*, 867460.

Pot-Kolder, R. M. C. A., Geraets, C. N. W., Veling, W., van Beilen, M., Staring, A. B. P., Gijsman, H. J., Delespaul, P. A. E. G., & van der Gaag, M. (2018). Virtual-reality-based cognitive behavioural therapy versus waiting list control for paranoid ideation and social avoidance in patients with psychotic disorders: a single-blind randomised controlled trial. *The Lancet Psychiatry*, *5*(3), 217–226.

Potts, C., Ennis, E., Bond, R. B., Mulvenna, M. D., McTear, M. F., Boyd, K., Broderick, T., Malcolm, M., Kuosmanen, L., & Nieminen, H. (2021). Chatbots to support mental wellbeing of people living in rural areas: can user groups contribute to co-design? *Journal of Technology in Behavioral Science*, *6*, 652–665.

Qi, Y. (2024). Pilot Quasi-Experimental Research on the Effectiveness of the Woebot AI Chatbot for Reducing Mild Depression Symptoms among Athletes. *International Journal of Human–Computer Interaction*, 1–8.

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, *151*, 102630.

Rendevski, N., Trajcevska, D., Dimovski, M., Veljanovski, K., Popov, A., Emini, N., & Veljanovski, D. (2022). Pc vr vs standalone vr fully-immersive applications: History, technical aspects and performance. *2022 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 1–4.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, *44*(10), 1827–1843.

Ruane, E., Farrell, S., & Ventresque, A. (2021). User perception of text-based chatbot personality. *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, 32–47.

Schiliro, F., Moustafa, N., & Beheshti, A. (2020). Cognitive privacy: AI-enabled privacy using EEG signals in the internet of things. *2020 Ieee 6th International Conference on Dependability in Sensor, Cloud and Big Data Systems and Application (Dependsys)*, 73–79.

Schulhoff, S., Pinto, J., Khan, A., Bouchard, L.-F., Si, C., Anati, S., Tagliabue, V., Kost, A., Carnahan, C., & Boyd-Graber, J. (2023). Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4945–4977.

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *ArXiv Preprint ArXiv:2308.03825*.

Silvervarg, A., Raukola, K., Haake, M., & Gulz, A. (2012). The effect of visual gender on abuse in conversation with ECAs. *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings 12*, 153–160.

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2023). Large language models (llms) and empathy-a systematic review. *MedRxiv*, 2023–2028.

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, *3*(1), 12.

Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious games: An overview*.

Sygel, K., & Wallinius, M. (2021). Immersive virtual reality simulation in forensic psychiatry and adjacent clinical fields: a review of current assessment and treatment methods for practitioners. *Frontiers in Psychiatry*, *12*, 673089.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *J Med Internet Res*, *22*(3), e16235. https://doi.org/10.2196/16235

Telles-Correia, D., Moreira, A. L., & Gonçalves, J. S. (2015). Hallucinations and related concepts—their conceptual background. *Frontiers in Psychology*, *6*, 991.

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., Adlam, K., Nordgreen, T., Bruvik, K. H., & Hovland, A. (2018). ICBT in routine care: a descriptive analysis of successful clinics in five countries. *Internet Interventions*, *13*, 108–115.

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V, Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456–464.

Valmaggia, L. R., Latif, L., Kempton, M. J., & Rus-Calafell, M. (2016). Virtual reality in the psychological treatment for mental health problems: An systematic review of recent evidence. *Psychiatry Research*, *236*, 189–195.

van Wezel, M. M. C., Croes, E. A. J., & Antheunis, M. L. (2021). "I'm Here for You": Can Social Chatbots Truly Support Their Users? A Literature Review. *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, 96–113.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

White, M. P., Yeo, N. L., Vassiljev, P., Lundstedt, R., Wallergård, M., Albin, M., & Lõhmus, M. (2018). A prescription for “nature”–the potential of using virtual nature in therapeutics. *Neuropsychiatric Disease and Treatment*, 3001–3013.

Wright, M., Getta, A. D., Green, A. O., Kickett, U. C., Kickett, A. H., McNamara, A. I., McNamara, U. A., Newman, A. M., Pell, A. C., & Penny, A. M. (2021). Co-designing health service evaluation tools that foreground first nation worldviews for better mental health and wellbeing outcomes. *International Journal of Environmental Research and Public Health*, *18*(16), 8555.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136.

ZEMČÍK, M. T. (2019). A brief history of chatbots. *DEStech Transactions on Computer Science and Engineering*, *10*.