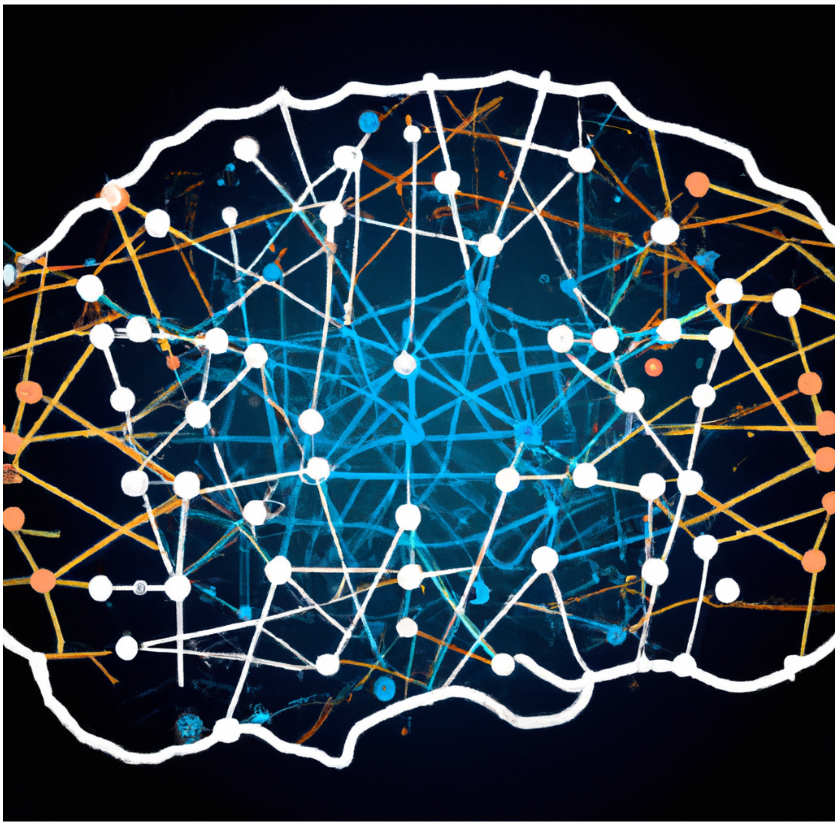

SEN2331

Master's Thesis

Design Guidelines for Integrating AI into Mental Healthcare: A Case Study on Posttraumatic Stress Disorder Prediction

Author :
T.P.S. Steurbaut

Student number :
5456282

2022-2023

Abstract

Global mental healthcare confronts daunting challenges, notably posttraumatic stress disorder (PTSD), necessitating immediate attention and innovative solutions. Many individuals requiring mental health support face various barriers, including social stigma, low perceived need, and restricted access to care providers — especially prevalent in certain regions — which impedes their quest for professional assistance. Amidst these obstacles, artificial intelligence (AI) emerges as a promising instrument, ushering in its unique challenges and opportunities. This thesis delves into these intricacies, aiming to develop comprehensive and pioneering design guidelines for AI applications within mental healthcare. The research focuses on three pivotal areas: identifying the distinct challenges and opportunities of implementing AI in mental healthcare; adapting existing AI design principles to fit the mental health landscape; and understanding the crucial role of multidisciplinary collaboration and user-centered design in this context. The primary objective is to devise guidelines that address inherent difficulties in mental healthcare, such as stigma and complexity of disorders, while leveraging potential benefits like early support intervention and expanded access to mental healthcare services. The suggested design guidelines embrace a systematic approach, encapsulating problem definition, stakeholder engagement, data acquisition, ethical and legal considerations, model design, system deployment, usage, maintenance, and iterative improvements based on feedback. Grounding these guidelines in a practical context, the thesis introduces AnchorAid, a tool that is designed theoretically and remains hypothetical. This virtual assistant provides post-trauma recovery support by gathering data, generating personalized feedback, recommending symptom management strategies, and assisting clinicians in the patient management processes. Through design guidelines establishment and their implementation via AnchorAid, this thesis lays a solid foundation for AI integration into mental healthcare.

Contents

1 Introduction

1.1 Background and Motivation

Mental healthcare, a critical facet of global health, confronts considerable challenges worldwide. Among these, posttraumatic stress disorder (PTSD), a severe psychiatric disorder, stands as both a pressing societal issue and a condition that individuals are often tasked with managing. The disorder's average lifetime prevalence is determined to be 7.34%, a statistic specifically derived from data collected in Australia, Canada, the Netherlands, New Zealand, and the USA (Dückers, Alisic, & Brewin, 2016). PTSD frequently manifests as recurrent distressing memories, dreams, dissociative reactions, and psychological distress. These symptoms can considerably compromise an individual's quality of life (Association, 2013). While the importance of professional intervention in managing PTSD is evident, several barriers including stigma, extended waiting times, and geographical limitations often discourage individuals from seeking necessary help (Andrade et al., 2014). Moreover, it's alarming to note that almost half of the individuals needing mental healthcare fail to engage with formal healthcare systems (Alonso et al., 2007). Such gaps in care provision emphasize the need for innovative solutions that can transcend these barriers, thereby enhancing the accessibility and efficacy of mental healthcare.

Artificial intelligence (AI), including machine learning (ML) and deep learning (DL), promises a new horizon of such innovative solutions. AI's proficiency in handling complex datasets and generating robust predictive models has proven indispensable across various healthcare domains, including oncology, dermatology, imaging, and now, mental health (Lee et al., 2021) (Shafiei, Lone, Elsayed, Hussein, & Guru, 2020). The continuous evolution of medical devices and digital record systems results in a vast amount of heterogeneous data, enabling the healthcare sector to significantly benefit from AI (Esteva et al., 2019). AI's potential extends to transforming mental healthcare delivery by predicting prognosis, designing personalized treatment plans, and more importantly, democratizing access to care. The synergistic relationship between AI and human clinicians can lead to enhanced care delivery, culminating in better patient outcomes (J. H. Chen & Asch, 2017).

However, challenges emerge with the integration of AI into mental healthcare. These encompass stigma, diagnostic and prognosis complexity, potential bias, limited access to expert knowledge, and subjectivity in mental healthcare (P. Corrigan, 2004). Yet, these challenges also present unique opportunities for AI to augment traditional methods, enhancing accessibility, facilitating early detection and intervention, minimizing reliance on self-reported information, and providing a more nuanced understanding of the complex causality of mental health disorders (Birnbaum, Rizvi, Correll, Kane, & Confino, 2017). In light of these opportunities and challenges, the Future of Life Institute (FLI) recently called for a pause on the training of AI systems more powerful than GPT-4, citing the urgent need to address risks related to opacity, biases, and misinformation (*Pause Giant AI Experiments: An Open Letter*, 2023). Multiple reports and regulations have also emphasized the importance of establishing safety protocols and guidelines for the responsible development and use of AI systems (*The Act*, 2021) (Tabassi, 2023) (U.S. Chamber of Commerce, 2023) (HM Government, 2021) (Chatila & Havens, 2019). Hence, the focus should be on implementing AI in mental healthcare using design guidelines that ensure seamless integration with the system and stakeholders (J. H. Chen & Asch, 2017).

1.2 Innovative Problem Statement and Research Questions

Despite the high lifetime prevalence of any potential traumatic event (PTE) that can reach up to 80% (de Vries & Olff, 2009), significant barriers persist in mental healthcare, leaving many individuals without necessary treatment (de Vries & Olff, 2009). The long delays before treatment initiation and high dropout rates from treatment programs highlight the severity of the situation (P. S. Wang, Berglund, Olfson, & Kessler, 2004) (Andrade et al., 2014). These issues underscore the pressing need for innovative technology, such as AI, to enhance the effectiveness of the mental healthcare system. However, as mentioned, the integration of AI into mental healthcare, while promising, is fraught with significant challenges. The focus should thus be on developing protocols, best practices, and guidelines to optimize the opportunities that AI presents while addressing the challenges. By leveraging AI's potential for early intervention, personalized support, enhanced decision-making, improved access to care, and advanced clinical research, the potential exists to revolutionize mental healthcare. In light of this, the primary goal of my thesis is to establish robust design guidelines for integrating AI into the mental healthcare sector. The research thus centers on the following question

- **How can we develop design guidelines to address challenges and leverage opportunities for AI systems in mental healthcare?**

To answer this, I will delve into the unique challenges and opportunities for AI in mental healthcare, investigate the adaptation of existing design principles for AI in general healthcare to mental healthcare, and explore how multidisciplinary collaboration and a user-centered design approach can contribute to the development and evaluation of AI systems in mental healthcare. This leads to the following sub-questions:

1. **What are the unique challenges and opportunities of AI in mental healthcare, and how do they shape the design requirements for AI systems?**
2. **How can existing design principles and guidelines for AI in general healthcare be adapted to the mental healthcare context?**
3. **How can multidisciplinary collaboration and a user-centered design approach contribute to the development and evaluation of AI systems in mental healthcare?**

I will critically analyze the current principles and guidelines for AI in general healthcare and evaluate their applicability to PTSD prognosis prediction case studies. Additionally, I will shed light on key considerations in the development of AI design guidelines for mental healthcare, including stigma, trust, the complexity of disorders, privacy concerns, and patient-centered care.

To illustrate the practical application of my research findings and the proposed guidelines, I will present a conceptual AI tool - AnchorAid, the virtual assistant. This tool, hypothetically designed to assist patients and care providers in managing PTSD, exemplifies the potential of AI in revolutionizing mental healthcare by providing more timely, personalized, and accessible care.

In summation, my thesis aspires to lead groundbreaking research into AI's integration into mental healthcare, especially focusing on PTSD, making it one of the first works in this field to combine these areas. The core objective is to formulate innovative and robust design guidelines for its integration into this sector. The unique approach of this exploration is exemplified by the conceptual tool, AnchorAid, detailed in subsection 6.2, which is hypothetically designed to enhance patient outcomes. This forward-looking investigation highlights the novelty of the research and its objectives, underpinned by a comprehensive literature review of the most up-to-date AI-related reports, regulations, and letters.

1.3 Alignment with CoSEM Program

My thesis closely embodies the Complex Systems Engineering and Management (CoSEM) Master's program's core tenets, with its comprehensive exploration of the integration of AI into the intricate socio-technical landscape of mental healthcare. This research appreciates the breadth of elements that come into play in such a complex system, echoing the CoSEM approach by taking into account the technical, social, regulatory, and ethical dimensions inherent in the design and implementation of AI systems in mental healthcare. The design guidelines proposed are envisioned to foster a seamless integration of AI into the existing mental healthcare system, addressing not just the technical aspects of the AI tool, but also the system of implementation, which includes regulatory concerns and user acceptance.

As the CoSEM program emphasizes, socio-technical environments are highly complex, requiring an understanding that extends beyond the scope of pure technology. Similarly, the integration of AI into mental healthcare involves an intricate interplay of multiple factors including care providers, patients, AI designers, and policymakers, making it an exemplar of a complex socio-technical system. My research thus navigates the complexities of stakeholder involvement, reflecting the holistic approach advocated by CoSEM.

Lastly, the AI tool I proposed, AnchorAid, for managing PTSD exemplifies the essence of the CoSEM program. While its technical aspects represent the innovation sought in the CoSEM program, the ethical implications of AI integration in mental healthcare highlight the need for a judicious approach to innovation. Critical considerations around privacy, bias, transparency, and consent are extensively addressed, mirroring CoSEM's emphasis on the ethical dimensions of socio-technical design.

1.4 Thesis Structure

To write this thesis, I adopt a structured and thorough approach to explore the integration of AI into mental healthcare, specifically focusing on PTSD prognosis prediction. The narrative unfolds progressively, beginning with an introduction that provides background, motivation, and problem statement, and outlines the research questions.

The methodology chapter elucidates the design science and grounded theory employed in the study. These approaches underpin the subsequent research and align with the semi-structured interviews used for data collection. A comprehensive literature review follows, discussing the challenges and opportunities for AI in mental healthcare and PTSD prognosis prediction. I examine existing design principles and guidelines for AI in healthcare, emphasizing the significance of multidisciplinary collaboration and human-centered design. An in-depth case study provides the context for the application of AI, offering a thorough understanding of PTSD, its prevalence, diagnosis, prognosis, and treatment. This section includes a review of current ML and DL methods relevant to PTSD prediction. With the groundwork laid, my thesis then presents the development of design guidelines for integrating AI into mental healthcare. This includes an analysis of the interview findings that informed these guidelines. The next chapter introduces the concept of virtual agents and unveils the culmination of my research – AnchorAid, a virtual support framework for post-trauma recovery designed in accordance with the newly developed guidelines. My thesis then transitions into a discussion and conclusion, reflecting on the implications, strengths, and limitations of the research, and hinting at potential future research directions. The structure ensures a logical progression that guides the reader from understanding the context and the problem, through the methodology and results, to the implications and conclusion. Supplementary materials incorporated into the appendices, including a glossary, and interview protocols, provide further support and depth to the research.

2 Pioneering Research: Novel Approach and Methods

2.1 Design Science Methodology

This paper focuses on the integration of AI into mental health settings, with the goal of improving the follow-up care of individuals with posttraumatic stress disorder. Although AI holds high potential for widespread real-world applications, its use in the medical field is still an emerging topic (Ochoa, Csiszár, & Schimper, 2021).

In this research, I developed an innovative approach that combines several methodologies. First, I employed design science research (DSR), a qualitative research methodology focused on the design process. This methodology generates knowledge about both the methods used to design an artifact and the artifact itself (Carstensen & Bernhard, 2019). Given that little research has been conducted on the concept of using AI to provide support to PTSD patients, I found the use of a qualitative approach beneficial and emphasized the importance of an exploratory phase to identify problem variables (Creswell, 2009). Furthermore, DSR aims to extend human and organisational capabilities by creating innovative artifacts, which aligns with my research topic (Hevner et al., 2004). The approach seems particularly effective when applied to innovative IT artifacts addressing information-related tasks to improve the functioning of organisations (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). Considering that AI systems work with large databases, this further justified my choice of this approach. I followed the three inherent research cycles of DSR, as shown in figure 1. The *relevance cycle* connects the research project's contextual environment to the design science activities by providing requirements and acceptance criteria for the evaluation of the results. The *rigor cycle* bridges the design science activities with the scientific foundations, expertise, and experience that guide the research project. The central *design cycle* repeatedly oscillates between the creation, evaluation, and refinement of the artifact (Hevner, 2007).

One strength of the design science research methodology is the continuous improvement cycle during which relevant information is gathered. Furthermore, this cycle addresses the changes that the artifact may effect not only within but also outside the organization in which the solution is implemented (Gregório et al., 2021). This ensures that the artifact fits the contextual and practical needs of society, politics, and law. However, limitations of this approach include the lack of cumulative development, perishable nature due to rapid advances in technology, and difficulty in applying rigorous evaluation methods (Hevner et al., 2004).

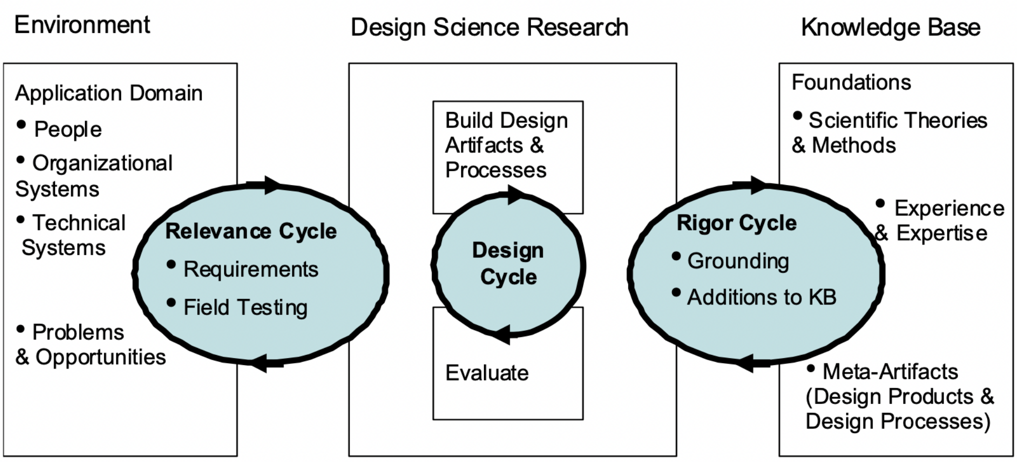

Figure 1: Design Science Research Cycles (Hevner, 2007)

Additionally, I integrated case study research (CSR) using interviews into my research process to facilitate ex-ante and ex-post evaluations of the artifact (Costa, Soares, & de Sousa, 2016). This approach proved to be useful for highlighting key aspects of healthcare delivery and gathering in-depth knowledge of the real-life context of PTSD patients' management (Crowe et al., 2011). I conducted and analyzed interviews based on grounded theory (GT), which is detailed in the next subsection 2.2.

2.2 Grounded Theory for Semi-structured Interviews

Methods based on grounded theories provide systematic yet flexible guidelines for gathering and analysing qualitative data to develop theories directly from the data (Charmaz, 2014). My aim was to draw meaningful conclusions based on issues concerning specific groups of people, identified with a data generation tool such as interviews. From inductive data, comparative methods, iterative strategies going back and forth between data and analysis, and continuous interaction between me, the data, and the analysis, I was able to construct theory. The interview questions needed to be general enough to include a variety of experiences while remaining narrow enough to capture each participant's unique experience (Gubrium, Holstein, Marvasti, & McKinney, 2012). I aimed at maintaining an equal position of power with the participant, achieved by letting the participant choose the scheduling and location of interviews, using a flexible approach to questioning, and being deeply involved in the research process. The meaning from the data was drawn in a mutual way, which required an open interchange. Memos described my thoughts on the research topic and how such perspectives might influence my analysis, to inform about the lens through which I looked at the data (Mills, Bonner, & Francis, 2006). It was crucial to maintain a clear connection between the participant experience analysis and the data from which it was derived. This method performed well when there was limited information available on the phenomenon, which was relevant in the case of deploying AI for mental healthcare (Chun Tie, Birks, & Francis, 2019). In addition, the delivery of healthcare is a dynamic process, influenced by a range of factors such as policy changes and the attitudes of healthcare practitioners. The use of grounded theory to identify and analyse these factors can provide valuable insights that can help practitioners improve their practices and better serve patients. It is important to note that healthcare is a collaborative effort involving a range of professionals, including clinicians, patients, managers, educationalists, and students. In order to promote mutual understanding and improve the quality of care, inter-professional education is increasingly being prioritised to address issues such as professional rivalries and ensure that all members of the healthcare team work together effectively (Immy, 2005).

2.2.1 Semi-structured Interviews

Sampling

Saturation is generally the guiding principle for sampling and is achieved when no new relevant data emerge from the interviews (Mason, 2010). However, for exploratory research such as mine, the goal isn't to cover all aspects of the area, but to provide new insights that significantly advance or contradict previous assumptions. My experienced research supervisor, Nadia Metoui, suggested that I interact with approximately ten participants during the exploratory phase. To fine-tune this estimation, I used a model evaluating five components determining the information power of the sample, which helped appraise the number of participants needed for the study (Malterud, Siersma, & Guassora, 2016). These five components are the aim of the study, sample specificity, use of established theory, quality of dialogue, and analysis strategy. Given the narrow focus of my study on AI, design guidelines, and a case study based on PTSD prognosis, this implies higher information power, hence a smaller sample size is needed. As I am working at the Douglas Mental Health University Institute, I utilized purposive sampling (Immy, 2005) by reaching psychologists, psychiatrists, and researchers in mental healthcare with knowledge specific to the area studied, which further reduced the required sample size. However, the topic of building design guidelines for using AI in mental healthcare lacks well-established theory, thereby increasing the number of informants required. Since some of the participants are part of the same research centre I work in, we can establish a good dialogue together and I require fewer participants. Finally, the analysis is case-specific, including an in-depth exploration of the discourse provided which increases information power and further reduces the sample size. Based on my research supervisor's advice, the analysis conducted above, and a similar case study where 13 participants were included (Andalibi & Flood, 2021), an initial estimation of 8 participants seemed feasible. However, this estimation was continuously refined throughout the research process, and ultimately, I conducted interviews with 11 participants. Indeed, continuous theoretical sampling helps fill gaps, clarify uncertainties, and further develop promising findings (Chun Tie, Birks, & Francis, 2019).

Data Collection

Interviews are the most common tool used to generate data in qualitative inquiry, and the semi-structured format is the most frequently used technique (Dicicco-Bloom & Crabtree, 2006), particularly in healthcare (Gill, Stewart, Treasure, & Chadwick, 2008). Semi-structured interviews are versatile, flexible, and allow for a reciprocal relationship between the researcher and the informant, with both parties at an equal position of power, aligning with grounded theory requirements (Immy, 2005). The questions are formulated in advance based on the interview guide (Mason, 2010), providing a structure that ensures coverage of the main research topics but does not need to be strictly followed. The goal is to guide participants to discuss the phenomenon while enabling the interviewer to improvise questions based on the informant's responses (Gill, Stewart, Treasure, & Chadwick, 2008). This semi-structured format is suitable for my research project because participants are unfamiliar with AI and aren't accustomed to discussing potential issues that could arise from it (Kallio, Pietilä, Johnson, & Kangasniemi, 2016). It also allows for a wide range of perceptions, essential in a complex socio-technical system like the one that supports mental healthcare (Cridland, Jones, Caputi, & Magee, 2015).

The following process is based on a framework for developing a qualitative semi-structured interview guide developed by Kallio, Pietilä, Johnson, & Kangasniemi (2016). After I conducted a comprehensive literature review, I formulated a preliminary semi-structured interview guide consisting of a list of questions directing the conversation towards the research topic. The questions needed to be participant-oriented, well-formulated, single-faceted, and open-ended. In particular, the words *what*, *who*, *where*, *when*, *how*, and *why* can motivate participants to provide descriptive answers. I first discussed the main themes where informants freely developed their perceptions about the subject, from lighter topics, moving to more in-depth ones, and then returning to lighter topics again (Adams, 2015). Predetermined or spontaneous follow-up questions helped maintain coherence between each participant's interview and extracted accurate information from their experience. It was valuable to include some questions that required short and straightforward answers to obtain quantitative data. I then tested my interview guide using the *expert assessment* technique by presenting it to my supervisors, who are accustomed to interviews for mental healthcare.

Data Analysis

The data was analysed following the constant comparative method, used in grounded theory, which supports a concurrent and iterative process of data generation and analysis (Immy, 2005). Coding helps make sense of the data by identifying concepts and similarities and categorising them through iterative phases. The framework developed by Chun Tie, Birks, & Francis (2019) provides descriptions of these phases. Regrettably, resource constraints precluded the verbatim transcription of the interviews, which, in turn, forestalled the execution of coding as originally intended. Nevertheless, I used memos and diagrams throughout the process to keep a record of my analysis, feelings, and intuitions for future data generation, and to visualise the data clearly and succinctly. 'Initial coding' classifies the data, gives it meaning, compares incident to incident, labels early tendencies, and starts looking for similarities between the codes. This helps make decisions about additional data collection. By selecting new individuals or materials that contain pertinent information, theoretical sampling enabled me to further explore patterns in the data. It was essential to ensure that the constructed theory was grounded in the data, which is the core of grounded theory. In the 'intermediate coding' phase, more abstract concepts can be derived from the basic data, which brings researchers closer to building a theory. Properties, the characteristics shared in a category, and dimensions, the variations of a property, are clarified.

It is possible to identify a core category at this stage, and reach theoretical saturation although this concept wasn't strictly applied in my research as explained in the paragraph addressing the sampling. During the 'advanced coding' phase, the storyline process creates a narrative that conceptualises the core category by linking the categories and yielding a group of theoretical propositions. Finally, theoretical coding reconstructs the broken story into a well-organised theory. Theoretical sensitivity represents the capacity to recognise when a data segment is significant to the construction of the theory and can be increased by reading literature, coding, memoing, and reflecting on the data throughout the entire research process.

2.3 An Innovative Approach to Research: Research Framework and Research Flow Diagram

The three main stages of my research, which combine DSR, CSR, and GT, adapted from a study on collaborative networks (Costa, Soares, & de Sousa, 2016), are shown in figure 2.

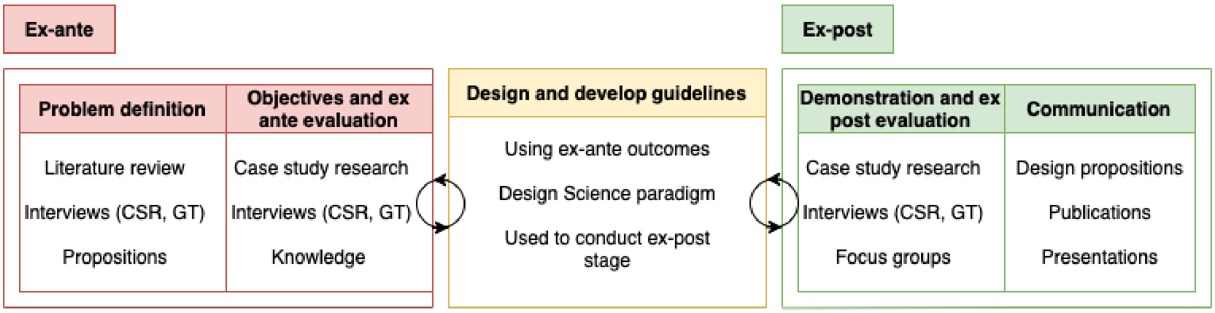

Figure 2: Research framework combining DSR, CSR, and GT

In line with the innovative approach that guided this research, I meticulously developed a comprehensive research flow diagram to illustrate the intricacies of the study. The research flow diagram, depicted in figure 3, outlines the three components of the DSR methodology - Environment, Design Science Research, and Knowledge Base - for each sub-question. The diagram also mentions the methods used to gather information, develop design guidelines, and assess new artifacts. The DSR methodology was integrated with CSR and GT, providing a relevant context for data collection, a holistic perspective, precise analysis, transferability, and reliability. The three cycles displayed for each sub-question represent the relevance cycle, design cycle, and rigor cycle of the DSR methodology. Iterating through these cycles and sub-questions enabled me to address the main research question, displayed at the bottom of the diagram.

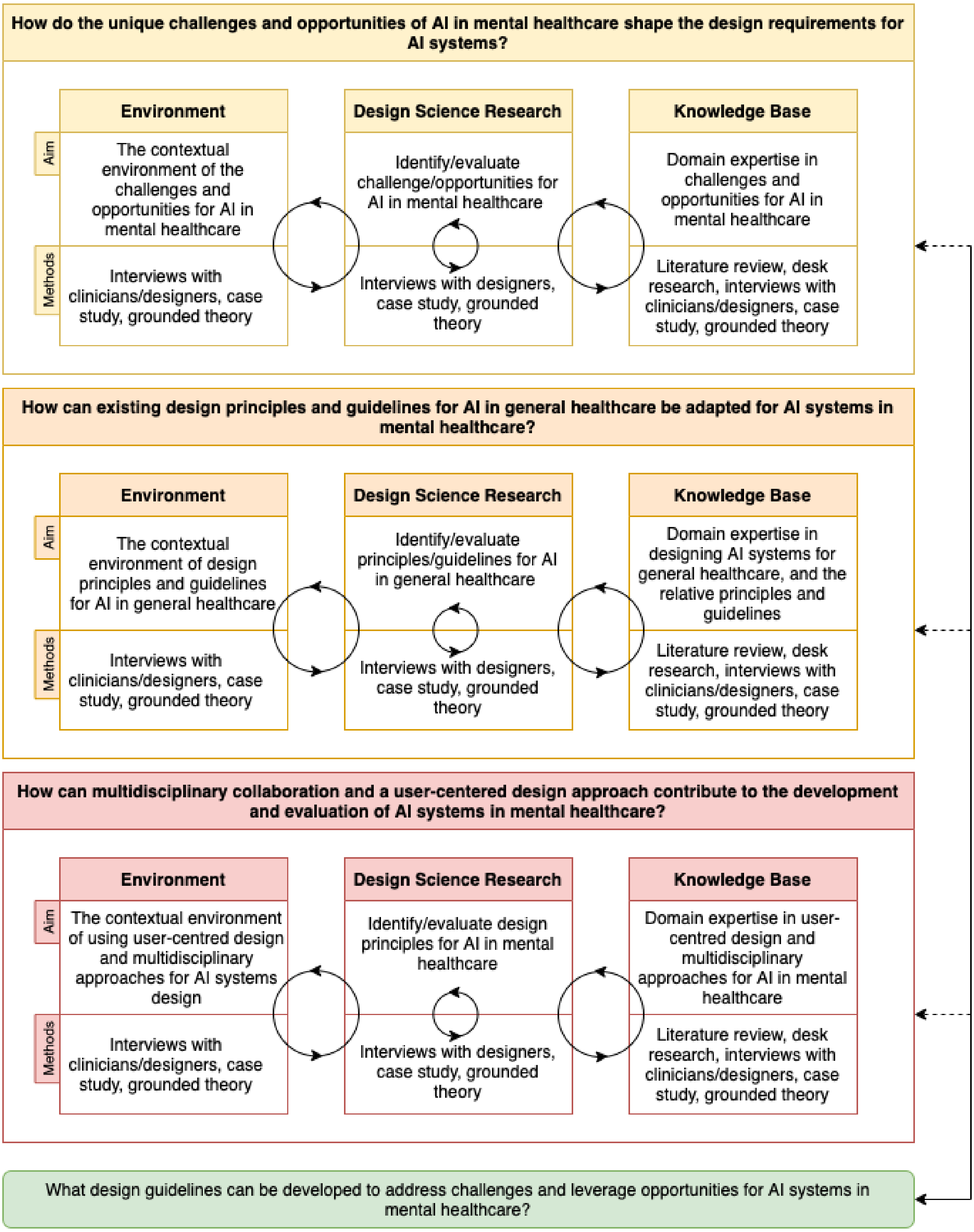

Figure 3: Research flow diagram

3 Literature Review

The goal of this literature review is to shed light on the challenges and opportunities in mental healthcare, particularly concerning the prediction of PTSD prognosis. Additionally, it aims to dissect the prevailing design principles and guidelines for AI applications within the broader healthcare sphere, while highlighting the importance of multidisciplinary collaboration and human-centered design. To ensure a thorough review, I performed a targeted search of scholarly articles on databases such as Scopus, ScienceDirect, and PubMed, employing specific keywords detailed in each subsection. To refine the search process, I used backward and forward snowballing methodologies, which helped in the careful selection of articles. The specific nature of the topic warranted the invocation of certain criteria to limit the scope of the results. Given the rapid evolution in the fields of mental health, anxiety-related disorders, and AI, I restricted the review to studies conducted after the year 2000. To maintain homogeneity and readability, articles not published in English were excluded from consideration.

3.1 Challenges and Opportunities in Mental Healthcare and Posttraumatic Stress Prognosis Prediction

Mental healthcare, especially the diagnosis, prognosis, and treatment of PTSD, offers a unique landscape of challenges and opportunities for incorporating AI technologies. In this section, I'll explore these crucial challenges and opportunities, with a particular focus on PTSD prognosis prediction. This analysis will guide the formation of effective design guidelines for AI systems in this domain.

Keywords: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Mental Health, Mental Healthcare, Posttraumatic Stress Disorder (PTSD), PTSD Prognosis Prediction, Healthcare Challenges, Healthcare Opportunities, Bias in Healthcare, Data Bias, Stigma in Mental Health

Subject of Study	Articles
Stigma and Healthcare Disparities	• P. Corrigan (2004) • Clement et al. (2015) • P. W. Corrigan, Druss, & Perlick (2014) • P. W. Corrigan & Rao (2012) • Fitzpatrick, Darcy, & Vierhile (2017) • Schueller, Hunter, Figueroa, & Aguilera (2019) • Cramer, Garcia-Gathright, Springer, & Reddy (2018)

Continued on next page

Table 1 – continued from previous page

Subject of Study	Articles
Complexity of Diagnosis and Prognosis	• Schultebraucks & Chang (2021) • Olbert, Gala, & Tupler (2014) • Armour, Fried, Deserno, Tsai, & Pietrzak (2017) • Kubota, Chen, & Little (2016) • Shatte, Hutchinson, & Teague (2019)
Limited Access to Expertise	• Wainberg et al. (2017) • Patel et al. (2018) • Barnett, Ray, Souza, & Mehrotra (2018) • D. Luxton, Pruitt, & Osenbach (2014)
Inherent Subjectivity and Reliance on Self-report	• Rosen & Lilienfeld (2008) • Dyer et al. (2009) • Onnela & Rauch (2016) • Gibbons et al. (2014)
Early Detection and Intervention	• Birnbaum, Rizvi, Correll, Kane, & Confino (2017) • Conway & O'Connor (2016) • Kazdin & Blase (2011) • Nehra et al. (2019) • Bohr & Memarzadeh (2020) • Naslund et al. (2017) • Mentis, Lee, & Roussos (2023)

Continued on next page

Table 1 – continued from previous page

Subject of Study	Articles
Scalable and Accessible Mental Healthcare	• Torous et al. (2018) • Firth et al. (2017) • Saxena, Sharan, Garrido, & Saraceno (2006) • Patel et al. (2016) • Ruggiero et al. (2015) • Hollis et al. (2017a)
Personalized and Adaptive Treatment	• Chekroud et al. (2016) • Mohr, Burns, Schueller, Clarke, & Klinkman (2013) • Ebert et al. (2018) • Bauer et al. (2020) • Campbell & Choudhury (2012) • Schultebraucks & Chang (2021) • Insel (2014) • Malgaroli & Schultebraucks (2020)
Enhanced Provider Decision-making	• Fakoor, Ladhak, Nazi, & Huber (2013) • Davenport & Kalakota (2019) • Luo, Wu, Gopukumar, & Zhao (2016) • Sittig & Singh (2010) • Bertl, Metsallik, & Ross (2022) • Philippe et al. (2022)

Table 1: Relevant articles addressing the challenges and opportunities in mental healthcare and posttraumatic stress prognosis prediction

3.1.1 Challenges in Mental Healthcare and PTSD Prognosis Prediction

1. **Stigma and Healthcare Disparities**: A significant challenge in mental healthcare is the stigma associated with mental health disorders, which can deter individuals from seek-

ing timely help (P. Corrigan, 2004) (Clement et al., 2015). It is a significant public health concern because it prevents people from seeking help and increase dropout rates of ongoing treatment. Public stigmas such as social judgment and rejection, employment discrimination, shame, and family stigma, contributes to the reluctance to access care (P. W. Corrigan, Druss, & Perlick, 2014). Once a person internalizes negative stereotypes, it can also result in self-imposed isolation (P. W. Corrigan & Rao, 2012). AI-driven tools, such as chatbots and digital platforms, can offer an anonymous and non-judgmental environment for individuals to discuss their mental health concerns, potentially reducing barriers to care and enhancing help-seeking behavior (Fitzpatrick, Darcy, & Vierhile, 2017). In particular, these tools can be beneficial for marginalized people who suffer from healthcare provision disparities (Schueller, Hunter, Figueroa, & Aguilera, 2019). However, AI systems frequently exhibit enhanced optimization for specific groups, while demonstrating suboptimal performance or even exclusion for others, often due to inherent biases. Consequently, a challenge is to design these systems with care to guarantee equitable access to care for all individuals, mitigating the impact of such biases (Cramer, Garcia-Gathright, Springer, & Reddy, 2018).

2. **Complexity of diagnosis and prognosis**: Mental health disorders, including PTSD, are often characterized by complex and heterogeneous symptom profiles, which can make accurate diagnosis and prognosis challenging. The complexity of clinical phenotypes in psychiatry including individual factors, social factors, and complex interactions hinders the use of AI methods for mental healthcare (Schultebraucks & Chang, 2021). The study conducted by Olbert, Gala, & Tupler (2014) states that two individuals meeting symptom criteria for PTSD share approximately 60% of symptoms in common. In some cases, two individuals can share very few or even no symptoms in common yet share the same diagnosis. Also, there are some symptoms that are commonly experienced by individuals with both Major Depressive Disorder (MDD) and PTSD (Olbert, Gala, & Tupler, 2014) (Armour, Fried, Deserno, Tsai, & Pietrzak, 2017). A challenge for AI technologies, such as machine learning (ML) and deep learning (DL) algorithms, and natural language processing, is to assist in improving the precision of diagnosis and prognosis. These models analyze vast amounts of patient data, including electronic health records, natural language, and data from wearable sensors, and identify predictors of symptom progression or treatment response that may not be readily discernible to human clinicians. For instance, AI can help to improve clinical diagnosis, prognosis, treatment, and research in Parkinson's and Alzheimer's diseases, depression, and schizophrenia (Kubota, Chen, & Little, 2016) (Shatte, Hutchinson, & Teague, 2019).

3. **Limited access to expertise**: Most people around the world continue to have limited and dispersed access to mental health treatments, in particular in low- and middle-income countries (LMICs). The scarcity of mental health professionals, particularly in rural and low-resource settings, can limit the availability and quality of care for individuals with PTSD and other mental health disorders (Wainberg et al., 2017). This deficiency often leads to a restricted pool of available care providers, exacerbating the problem by creating lengthy waiting lists for those seeking help. Digital technology has the ability to substantially change mental healthcare, particularly through educating and assisting clinicians, monitoring treatment procedures, and encouraging self-help. It should not be utilized as a replacement for traditional approaches to mental health therapy, but rather as an additional tool (Patel et al., 2018). AI-driven tools, such as telepsychiatry platforms, clinical decision support systems, and intelligent virtual agents, can help bridge the gap between the demand and supply of mental health services by extending the reach of expert care (Barnett, Ray, Souza, & Mehrotra, 2018). Digital technology enables practitioners to efficiently track symptoms and other health-related factors between face-to-face or telehealth

treatment sessions. It facilitates the provision of tailored services to meet the specific needs of each patient. Furthermore, it can improve access to services for patients who speak various languages by connecting them with clinicians from different geographical locations. Intelligent virtual agents have the ability to engage in social interactions that resemble human behavior. These agents can be created to perform clinical interviews, assess the results, and provide feedback to patients (D. Luxton, Pruitt, & Osenbach, 2014). Therefore, the challenge for AI in mental healthcare is not just to develop effective tools, but to enhance general access to these resources around the world, particularly in lower-middle-income countries.

4. **Inherent subjectivity and reliance on self-report**: The process of diagnosing, establishing prognosis, and treating PTSD often hinges on patients' self-reported symptoms. However, these self-reports can be influenced by factors such as recall bias, where patients may not accurately remember past experiences or symptoms, and social desirability bias, where patients may underreport or overreport symptoms to align with perceived social expectations (Rosen & Lilienfeld, 2008) (Dyer et al., 2009). These biases can affect the reliability and validity of the collected data, ultimately impacting the quality of care. AI technologies offer the potential to supplement self-report assessments with more objective and reliable measures of patients' mental health status. Wearable sensors, passive data collection from smartphones, and virtual agents can provide real-time, continuous monitoring of patients' symptoms and behaviors. They can potentially enhance the accuracy and timeliness of care, as well as reduce the assessment burden on both patients and clinicians, by addressing the challenges of maintaining quality, relevance, and continuity in the data (Onnela & Rauch, 2016) (Gibbons et al., 2014).

3.1.2 Opportunities in Mental Healthcare and PTSD Prognosis Prediction

1. **Early detection and intervention:** AI technologies can facilitate the early detection and intervention of PTSD and other mental health disorders by analyzing large-scale data from electronic health records, social media, or other sources to identify at-risk individuals and provide targeted preventive interventions (Birnbaum, Rizvi, Correll, Kane, & Confino, 2017) (Conway & O'Connor, 2016). Early intervention can improve patient outcomes, reduce the burden on healthcare systems, and enhance the overall cost-effectiveness of mental health services (Kazdin & Blase, 2011) (Nehra et al., 2019). Furthermore, AI-driven tools can facilitate continuous monitoring of individuals exposed to traumatic events, enabling timely identification of those who may benefit from early intervention, and potentially preventing the development or exacerbation of PTSD symptoms. For instance, an MIT-based lab has studied the use of neural networks to identify early depressive symptoms in speech (Bohr & Memarzadeh, 2020) (Naslund et al., 2017). In addition, early detection and management of chronic stress play a vital role in preventing several diseases, not only PTSD. By leveraging AI, improved outcomes can be achieved through timely detection and intervention (Mentis, Lee, & Roussos, 2023).

2. **Scalable and accessible mental healthcare:** AI technologies can enable the development of scalable and accessible mental healthcare solutions, such as digital therapeutics, mobile apps, and chatbots. Mental health apps exist for depression, anxiety, and bipolar disorders and can provide evidence-based interventions to a wide range of users, regardless of their geographic location, socioeconomic status, or access to professional care (Torous et al., 2018) (Firth et al., 2017). Such solutions can help address the global mental health treatment gap and democratize access to quality care, particularly for underserved populations. Indeed, a majority of low-income countries and a significant portion of lower middle-income countries have less than one psychiatrist per 100,000 people, highlighting a shortage of mental health professionals (Saxena, Sharan, Garrido, & Saraceno, 2006)

(Patel et al., 2016). Web-based interventions have shown great benefits relative to PTSD symptoms and access to clinical expertise but this area is still under-researched (Ruggiero et al., 2015) (Hollis et al., 2017b).

3. **Personalized and adaptive treatment:** AI-driven tools can support the development and delivery of personalized and adaptive treatments for PTSD and other mental health disorders. They can continuously monitor patient's symptoms, treatment adherence, and response to interventions, and adjust treatment plans based on individual needs and preferences (Chekroud et al., 2016). This approach can enhance the effectiveness and efficiency of care, as well as patient engagement and satisfaction. Indeed, the interventions are tailored to individual patients' characteristics, such as their genetic, neurobiological, or psychological profiles, and provide real-time feedback and support (Mohr, Burns, Schueller, Clarke, & Klinkman, 2013) (Ebert et al., 2018). Harnessing AI techniques is a promising and essential avenue for advancing precision psychiatry, as the wide-ranging variations in the presentation and underlying causes of PTSD result in unique characteristics among patients. Integrating data from diverse sources into AI technologies is crucial (Bauer et al., 2020) (Campbell & Choudhury, 2012), in particular for PTSD management. It allows identifying subgroups of patients with distinct clinical profiles, treatment preferences, or risk factors. This integration can help overcome the limitations of traditional self-report assessments and enable more precise and personalized care for individuals with PTSD and other mental health disorders (Schultebraucks & Chang, 2021). AI can thus facilitate the development of personalized treatments and lead to precision psychiatry, but this requires careful handling and coordination (Malgaroli & Schultebraucks, 2020) (Insel, 2014).

4. **Enhanced provider decision-making:** AI-driven decision support systems can assist mental healthcare providers in making more informed clinical decisions by synthesizing large volumes of patient data, such as electronic health records, genomics, and real-time monitoring data. For instance, neural networks have been used to recognize potentially cancerous lesions in radiology images (Fakoor, Ladhak, Nazi, & Huber, 2013). Natural Language Processing (NLP) systems can analyze unstructured clinical notes and conduct conversational AI with patients (Davenport & Kalakota, 2019). By harnessing big data and AI technologies, clinicians are empowered to gain a comprehensive understanding of a patient's health condition, facilitating enhanced data collection and analysis with improved accuracy (Luo, Wu, Gopukumar, & Zhao, 2016). This can lead to improved clinical outcomes, reduced treatment variability, and more efficient use of clinical content in mental healthcare, as well as better collaboration and communication of care among multidisciplinary teams (Sittig & Singh, 2010). However, the existing research lacks comprehensive and consistent findings. There is still a gap between the technical possibilities of digital decision support systems for PTSD and real-world clinical work (Bertl, Metsallik, & Ross, 2022). Establishing new guidelines for reporting, evaluating, and assessing quality and bias would expedite literature synthesis and enhance confidence in provider decision-making (Philippe et al., 2022).

In summary, the primary hurdle for AI in healthcare lies not in the capabilities of the technologies themselves, but rather in ensuring their integration and acceptance into everyday clinical practice. The unique challenges and opportunities in mental healthcare and PTSD prognosis prediction shape the design requirements for AI applications in this domain. By addressing these challenges and leveraging the opportunities, AI technologies can contribute to more effective, efficient, and equitable mental healthcare services and outcomes for individuals with PTSD and other mental health disorders. The insights gained from this literature review will inform the development of design guidelines for AI integration in mental healthcare, as discussed in the following sections.

3.2 Existing Design Principles and Guidelines for AI in General Healthcare

Numerous organizations, governments, and researchers have proposed design principles, guidelines, and best practices for AI systems in healthcare. To synthesize these multifaceted insights, I've systematically reviewed, compared, and integrated these proposals. The resulting summary organizes these guiding principles into distinct categories. In addition, this section explores how these principles can be adapted to create effective design guidelines for AI applications in mental healthcare, particularly in the context of PTSD prognosis prediction. The focus is on understanding how these guidelines address the challenges and leverage the opportunities for AI in mental healthcare, identified in subsection 3.1.

Keywords: Guidelines, Best Practices, Design Principles, Ethical Considerations, AI, Machine Learning, Healthcare AI, AI in Healthcare, Mental Healthcare, AI Ethics in Healthcare

Subject of Study	**Articles**
Fairness and accountability	• *The Act* (2021) • Tabassi (2023) • U.S. Chamber of Commerce (2023) • HM Government (2021) • Weiser (2019) • Cramer, Garcia-Gathright, Springer, & Reddy (2018) • Amershi et al. (2019) • I. Y. Chen, Szolovits, & Ghassemi (2019) • Shin (2020) • Chatila & Havens (2019)

Continued on next page

Table 2 – continued from previous page

Subject of Study	Articles
Transparency, interpretability, and explainability	• *The Act* (2021) • Tabassi (2023) • U.S. Chamber of Commerce (2023) • HM Government (2021) • Chatila & Havens (2019) • Weiser (2019) • Joyce, Kormilitzin, Smith, & Cipriani (2023) • Antoniou, Papadakis, & Baryannis (2022) • Amershi et al. (2019) • Shin, Zhong, & Biocca (2020)
Clinical validation and efficacy	• *The Act* (2021) • U.S. Chamber of Commerce (2023) • Topol (2019) • Mutasa, Sun, & Ha (2020) • Keane & Topol (2018) • Park, Choi, & Byeon (2021) • Tabassi (2023) • Anhang Price et al. (2014) • Amershi et al. (2019)

Continued on next page

Table 2 – continued from previous page

Subject of Study	Articles
Privacy and security	• *The Act* (2021) • Tabassi (2023) • U.S. Chamber of Commerce (2023) • HM Government (2021) • Weiser (2019) • Branch (2019) • *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)* (2016) • Rep. Pallone (2022) • Forcier, Gallois, Mullan, & Joly (2019) • Abouelmehdi, Beni-Hessane, & Khaloufi (2018)

Continued on next page

Table 2 – continued from previous page

Subject of Study	Articles
human-centered design and interoperability	• *The Act* (2021) • Tabassi (2023) • De Silva & Alahakoon (2022) • U.S. Chamber of Commerce (2023) • HM Government (2021) • Weiser (2019) • Jin, Carpendale, Fraser, Hamarneh, & Gromala (2019) • Bates, Auerbach, Schulam, Wright, & Saria (2020) • Schueller, Hunter, Figueroa, & Aguilera (2019) • Amershi et al. (2019) • Mandl & Kohane (2012) • Patel et al. (2018)

Table 2: Relevant articles addressing existing design principles and guidelines for AI in general healthcare

3.2.1 Design Principles and Guidelines for AI in General Healthcare

1. **Fairness and accountability:** AI systems in healthcare should aim to promote fairness, manage harmful bias, and ensure accountability for system performance (*The Act*, 2021) (Tabassi, 2023) (U.S. Chamber of Commerce, 2023) (HM Government, 2021). If fairness is not meticulously addressed throughout the entire AI design process, the system runs the risk of generating discriminatory outcomes that disproportionately affect specific groups or individuals (Weiser, 2019) (Cramer, Garcia-Gathright, Springer, & Reddy, 2018). The design should match relevant social norms and mitigate social bias (Amershi et al., 2019). This is particularly significant in mental healthcare, as AI-driven tools can potentially perpetuate existing disparities and stigmatize specific populations (I. Y. Chen, Szolovits, & Ghassemi, 2019). The fairness of an algorithm can be evaluated by considering its accuracy (the rate of correct predictions), recall (the ability to identify relevant outcomes), and precision (the ability to provide accurate results) (Shin, 2020). Furthermore, AI systems designers and operators must be responsible and accountable when such systems cause harm (Weiser, 2019). Given that algorithms cannot be held legally responsible, it is imperative to ensure human accountability is integrated into every phase of content generation. This means that individuals involved in the process must bear responsibility for the outcomes and implications of the AI systems, emphasizing the need for human oversight and ethical considerations (Chatila & Havens, 2019) (Shin, 2020).

2. **Transparency, interpretability, and explainability:** To ensure that users comprehend the purposes and potential impact of an AI system, it is necessary for the system to be explainable and interpretable (*The Act*, 2021) (Tabassi, 2023) (U.S. Chamber of Commerce, 2023). It requires the system to be transparent, where the data and rules incorporated are made accessible for the purpose of testing and assessing potential risks (HM Government, 2021) (Chatila & Havens, 2019). Transparency ensures that the feature space, data sources, and algorithmic processes are well-documented and can be scrutinized for their appropriateness and relevance to the clinical problem and patient population at hand (Weiser, 2019) (Joyce, Kormilitzin, Smith, & Cipriani, 2023). Interpretability, on the other hand, focuses on the meaningfulness of the relationships between input features and the model's predictions (Antoniou, Papadakis, & Baryannis, 2022), allowing healthcare professionals to understand and trust the system's recommendations (Joyce, Kormilitzin, Smith, & Cipriani, 2023). Therefore, it must be clear what the system can do and how well the system can do it (Amershi et al., 2019). The concept of explainability bridges the gap between transparency and interpretability by providing a clear, human-understandable rationale for a model's prediction mechanisms in the healthcare domain (Joyce, Kormilitzin, Smith, & Cipriani, 2023). It aims to make the inner workings of AI systems more accessible to non-experts, including healthcare professionals and patients, allowing them to understand not only the input-output relationships but also the reasoning behind the model's decisions. The design should make clear why the system did what it did by providing explanations to the user (Amershi et al., 2019). Explainability helps to build trust in the system and promotes the adoption of AI-based solutions in healthcare, where understanding the rationale behind decisions is critical for patient safety, ethical considerations, and informed decision-making (Joyce, Kormilitzin, Smith, & Cipriani, 2023). In psychiatry, the need for explainability is crucial due to the probabilistic relationships between data describing syndromes, outcomes, disorders, and signs/symptoms, as well as the complex and multifaceted nature of the underlying causes (Antoniou, Papadakis, & Baryannis, 2022). When individuals have knowledge of how an algorithm operates and how its decisions are made, they are more inclined to use the content correctly and place trust in both the algorithm and the resulting outputs (Shin, Zhong, & Biocca, 2020).

3. **Clinical validation and efficacy:** AI systems must undergo rigorous testing and validation for clinical efficacy and safety before deployment in real-world settings (*The Act*, 2021) (U.S. Chamber of Commerce, 2023). The validation of the performance of a model does not always imply clinical efficacy, this is referred to as the 'AI chasm'. An Area Under the Curve (AUC), representing the accuracy of the algorithm, of 0.99 can still be worth nothing if clinical outcomes are not improved (Topol, 2019). Developing an algorithm that performs effectively on a small dataset from a specific population is different from creating one that can be applied to various populations and medical devices. Hence, AI models often demonstrate high accuracy when trained on specific data, but their performance tends to decline when presented with new, unseen data. This phenomenon is commonly known as 'overfitting' (Mutasa, Sun, & Ha, 2020). Furthermore, there is a significant disparity between the experimental code produced for a proof-of-concept research study and the final code intended for a regulated product. The latter is considered a medical device, necessitating a complete rewrite, implementation of a quality management system, and adherence to Good Manufacturing Practice standards (Keane & Topol, 2018). Clinical validation of AI algorithms for medical diagnosis and prediction requires a comprehensive evaluation process. This includes both internal and external validation. Internal validation involves techniques like cross-validation, where the original data is divided into testing and training sets, or split-sample validation, where the data is split into training, tuning, and test sets (Park, Choi, & Byeon, 2021). External validation is crucial and involves independent data collection to assess the algorithm's performance accurately. Diagnostic

case-control and cohort studies are used for this purpose, evaluating technical and clinical performance in real-world clinical scenarios (Park, Choi, & Byeon, 2021). In addition, assessing the risks associated with AI in a laboratory environment can provide valuable insights before deployment, but the risks that arise in real-world settings may vary from those initially measured (Tabassi, 2023). Lastly, to determine clinical utility, randomized clinical trials are conducted to measure the impact on patient outcomes. This evaluation cycle, which considers patient care experiences, is vital to demonstrate the effectiveness of AI tools in improving patient outcomes and informing clinical decision-making in mental healthcare (Park, Choi, & Byeon, 2021) (Anhang Price et al., 2014). Also, the AI systems should be designed to remember recent interactions, learn from user behavior and actions, and update based on user feedback (Amershi et al., 2019). Given that algorithms can evolve through continuous learning, changes must be notified to users and evaluation must be ongoing (Amershi et al., 2019).

4. **Privacy and security:** When designing, developing, and deploying AI systems, it is crucial to prioritize privacy values, including anonymity, confidentiality, and control, as these guide the system's overall approach to handling personal data (*The Act*, 2021) (Tabassi, 2023). Additionally, security is also of paramount importance and includes not only resilience, but also the implementation of protocols that prevent, protect against, respond to, and recover from potential attacks on the system (*The Act*, 2021) (Tabassi, 2023) (U.S. Chamber of Commerce, 2023) (HM Government, 2021). AI systems must be technically robust and safe. This includes demonstrating resilience to potential attacks, adhering to established safety protocols, and having a reliable fall-back plan in case of failure (Weiser, 2019). Such systems in healthcare must adhere to data privacy and security regulations, such as the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada (Branch, 2019), the General Data Protection Regulation (GDPR) in the European Union (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, 2016), and the proposed American Data Privacy and Protection Act (ADPPA) in the United States (Rep. Pallone, 2022) (U.S. Chamber of Commerce, 2023) (Forcier, Gallois, Mullan, & Joly, 2019). Safeguarding patients' sensitive mental health information is crucial for maintaining trust and ensuring the ethical use of AI technologies in mental healthcare. By complying with appropriate regulations (U.S. Chamber of Commerce, 2023) and implementing robust data protection measures, healthcare providers and AI developers can preserve patient confidentiality and minimize the risk of unauthorized access, data breaches, or other potential threats to sensitive information (Abouelmehdi, Beni-Hessane, & Khaloufi, 2018).

5. **Human-centered design and interoperability:** To ensure that AI systems are user-friendly and effective, it is important to incorporate HCD practices throughout the AI lifecycle (*The Act*, 2021) (Tabassi, 2023) (De Silva & Alahakoon, 2022). This includes promoting the active involvement of end-users and other relevant stakeholders, evaluating and adapting the end-user experience, and integrating human dynamics in all phases of the design process. Indeed, AI systems should be designed with the needs, preferences, and capabilities of end-users in mind, including patients, clinicians, and other stakeholders in the healthcare ecosystem (Tabassi, 2023) (U.S. Chamber of Commerce, 2023) (HM Government, 2021). To achieve this, AI systems should act as enablers of a democratic and equitable society by supporting user agency and fundamental rights. Additionally, human oversight should be allowed to maintain accountability and ensure the ethical use of the technology (Weiser, 2019). In the medical field, it requires getting clinicians, patients, and ethicists engaged in the design process (Jin, Carpendale, Fraser, Hamarneh, & Gromala,

2019) (Bates, Auerbach, Schulam, Wright, & Saria, 2020). In mental healthcare, designing for humans entails considering factors such as accessibility, usability, bias, and the potential impact of AI-driven interventions on mental health symptoms. It necessitates customized interventions designed specifically to address the requirements of the target population (Schueller, Hunter, Figueroa, & Aguilera, 2019). The design should ensure that the task the AI system performs is clear and responds to the end-users needs, and the information it provides is clinically relevant, considering context and potential biases (Amershi et al., 2019). In addition, AI applications must be designed to integrate seamlessly with existing healthcare systems, workflows, and data sources to minimize disruption and ensure continuity of care (Mandl & Kohane, 2012). The design should allow interaction with clinicians to support easy invocation, dismissal, or correction of AI systems output. It should scope down services when the system is uncertain about a user's goal (Amershi et al., 2019). Interoperability is especially important in mental healthcare, where fragmented and siloed care is a common challenge. Indeed, there is a lack of coherent plans to address mental health issues (Patel et al., 2018).

3.2.2 Compliance of Recent PTSD Prediction Studies with Design Principles for AI in General Healthcare

With the aim of designing guidelines specifically tailored for mental healthcare, I conducted an analysis of three recent articles focusing on PTSD prognosis prediction. This analysis was performed through the lens of the challenges and opportunities outlined in section 3.1, as well as the existing guidelines for general healthcare discussed in section 3.2.1.

1. 'Predicting Posttraumatic Stress Disorder Risk: A Machine Learning Approach' by Wshah, Skalka, & Price (2019):

   Early intervention has emerged as a promising approach to mitigate the long-term effects of PTSD and thus enhance **clinical efficacy** (Kearns, Ressler, Zatzick, & Rothbaum, 2012). However, due to the high costs of healthcare, providing early intervention to every individual who has experienced a traumatic event is not feasible. Interestingly, data collected within days following a traumatic event already exhibit predictive qualities for forecasting the prognosis of PTSD (I. R. Galatzer-Levy, Karstoft, Statnikov, & Shalev, 2014). During this critical period, AI systems can assist in monitoring individuals' mental health status through various means such as social media analysis, smartphone surveys, and interactions with virtual agents (Onnela & Rauch, 2016). The study conducted by (Wshah, Skalka, & Price, 2019) demonstrated an AUC of 0.85 for the prediction of high-severity PTSD symptoms. They used self-reported symptoms from data collected 1 month after trauma via smartphones to predict PTSD symptoms evolution and identify individuals who are at risk. A limitation of this approach is that self-reported symptoms can be influenced by biases and recall limitations (Rosen & Lilienfeld, 2008). Supplementing self-report assessments with more objective and reliable measures of mental health status can help capture more subtle patterns within the data (Kubota, Chen, & Little, 2016). AI technologies can leverage wearable sensors, passive data collection from smartphones, and virtual agents to provide real-time and continuous monitoring of symptoms and behaviors, improving the accuracy and timeliness of care (Onnela & Rauch, 2016). They highlighted the importance of proper feature selection to reduce overfitting, improve generalization, and increases accuracy (Wshah, Skalka, & Price, 2019) (Mutasa, Sun, & Ha, 2020). They also mentioned as further work the inclusion of additional data such as demographic variables and trauma histories which can help better capture the complexity of PTSD Wshah, Skalka, & Price (2019). Mental health disorders exhibit complex and heterogeneous symptom profiles, making accurate diagnosis and prognosis challenging (Schultebraucks & Chang, 2021). However, although they evaluated the performance of their model through cross-validation, they did not integrate external validation and randomized clinical trials to ensure clinical performance and utility in real-world clinical scenarios (Wshah, Skalka, & Price, 2019) (Park, Choi, & Byeon, 2021).

In addition, the participants (N=90) were predominately white (n=80) and males (n=57) Wshah, Skalka, & Price (2019). It is crucial to emphasize **fairness** and promote accessibility and equity in mental healthcare. AI systems should be designed to address barriers to care, such as geographical, financial, or cultural obstacles, and to support underserved populations (Torous et al., 2018) (Patel et al., 2018). Designers should also be mindful of potential digital divides and work to create inclusive technologies that cater to the diverse needs of patients, including those with limited digital literacy or access to technology.

Third, the study lacks clarification on how to communicate the rationale behind the model's decisions to clinicians Wshah, Skalka, & Price (2019) (Amershi et al., 2019). Addressing the issue of **explainability** is essential for future studies focusing on the diagnosis, prognosis, or treatment of mental health disorders to facilitate the adoption of such tools in clinical practice (Joyce, Kormilitzin, Smith, & Cipriani, 2023).

Furthermore, incorporating principles of **human-centered design** can contribute to the development of AI systems that align with clinicians' needs and seamlessly integrate into existing healthcare workflows, thereby promoting their adoption (Mandl & Kohane, 2012) (*The Act*, 2021) (Tabassi, 2023) (U.S. Chamber of Commerce, 2023) (HM Government, 2021) (Weiser, 2019). Moreover, human-centered design can facilitate the creation of AI systems tailored to support shared decision-making, empower patients in managing their conditions, and provide personalized recommendations based on individual circumstances (Malgaroli & Schultebraucks, 2020; Schueller, Hunter, Figueroa, & Aguilera, 2019). These systems can then adjust treatment plans based on individual needs and preferences, improving effectiveness and patient engagement (Mohr, Burns, Schueller, Clarke, & Klinkman, 2013) (Insel, 2014). Therefore, to address the complex and multifaceted nature of mental health issues, collaboration among different disciplines such as psychiatry, psychology, social work, and primary care is crucial (Patel et al., 2018; Bates, Auerbach, Schulam, Wright, & Saria, 2020).

Lastly, it is vital for further studies to mention the implementation of robust data protection measures, including encryption and anonymization, along with the establishment of strict protocols for data access and sharing. Protecting the **privacy and security** of patients' sensitive mental health data is a fundamental component in developing a safe and effective AI system for PTSD prognosis prediction (Abouelmehdi, Beni-Hessane, & Khaloufi, 2018) (*The Act*, 2021) (Weiser, 2019).

2. 'Re-examining Cortisol's Role in Predicting Post-traumatic Stress: Utilizing Machine Learning to Assess the Development of Non-remitting PTSD' by I. R. Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev (2017):

   This study employed supervised and unsupervised ML techniques to investigate the relationships between background, environmental, and neuroendocrine risk factors and the development of PTSD, taking into account the disorder's heterogeneity. They achieved stable predictive accuracy to identify trajectories of posttraumatic symptoms (AUC=0.82) The authors underscore the advantages of early prediction and intervention for **clinical efficacy** in preventing PTSD development. Nonetheless, incorporating passive data collection, wearable sensors, and virtual agents to supplement the emergency room (ER) evaluations and interviews they conducted could more effectively address biases and recall limitations (Onnela & Rauch, 2016). They acknowledge the necessity for further validation using independent samples, but this was not performed in the study (I. R. Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev, 2017).

Fairness was not explicitly addressed, as the information about participant demographics in the sample is limited. The authors mention the inclusion of gender, education level, and income data, but do not provide the corresponding proportions (I. R. Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev, 2017).

The results hold significant promise for clinical research and practice. The analysis revealed that lower cortisol levels are associated with a higher risk of PTSD non-remission, particularly for individuals who experienced early childhood trauma. Cortisol-based interventions can be specifically targeted to this group of individuals (I. R. Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev, 2017). Although I acknowledge that this study aims to uncover clinically significant features in heterogeneous samples rather than developing a tool directly for clinical usage, the **explainability** of the model and its predictions are not discussed in the article. If clinicians desire to use in the future a similar tool to identify at-risk individuals or target patients potentially responsive to a particular treatment, they will require a certain degree of trust in the model. This does not imply that they must comprehend every technical aspect of the model's predictions, but they should be able to interact with a transparent AI system that facilitates easy invocation, dismissal, or correction of errors when they occur (Amershi et al., 2019).

3. 'A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor' by Schultebraucks et al. (2020):

   In this study, an algorithm was developed to predict the posttraumatic stress course over 12 months upon ED admission after trauma. Indeed, trauma survivors often have their first and sometimes only interaction with the healthcare system during an emergency department (ED) visit. Interestingly, they mentioned that the limited external validation in existing studies makes it difficult to assess the generalizability of these models' performance and hinders the practical implementation of such algorithms in clinical practice (Schultebraucks et al., 2020). Therefore, they used an external validation sample. Among patients identified by the algorithm as likely to experience persistent posttraumatic stress (PTS) symptoms over a 12 months, 90% indeed exhibited such symptoms (Schultebraucks et al., 2020). This promising algorithm has the potential to enhance early clinical screening for long-term PTSD risk in the population of ED patients who have undergone a traumatic event. Its implementation could significantly improve **clinical efficacy** in this context.

   However, it is important to further investigate the issue of **fairness** to ensure the validity of the predictive algorithm in diverse clinical settings. The data used for model development were collected from patients admitted to the emergency departments (ED) of two specific sites in the United States, focusing on particular types of trauma. One limitation in the study was that the validation sample consisted of only 37.1% female patients. Schultebraucks et al. (2020) highlighted the need to evaluate the ecological validity of the algorithm in more heterogeneous ED populations (Schultebraucks et al., 2020).

   To address the concern of **explainability**, the researchers implemented methods to ensure the transparency of their models. They utilized SHAP (SHapley Additive exPlanations) feature importance ranking to identify which features have the most significant predictive power in distinguishing between non-remitting and resilient PTSD trajectories (Schultebraucks et al., 2020). This approach aims to provide insights into the factors that contribute most to the predictive outcomes of the algorithm.

Article	Design Principles
Wshah, Skalka, & Price (2019)	• Fairness and Accountability ✗ • Transparency, Interpretability, and Explainability ✗ • Clinical Validation and Efficacy ≈ • Privacy and Security ✗ • Human-centered Design and Interoperability ✗
I. R. Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev (2017)	• Fairness and Accountability ≈ • Transparency, Interpretability, and Explainability ✗ • Clinical Validation and Efficacy ≈ • Privacy and Security ✗ • Human-centered Design and Interoperability ✗
Schultebraucks et al. (2020)	• Fairness and Accountability ≈ • Transparency, Interpretability, and Explainability ✓ • Clinical Validation and Efficacy ✓ • Privacy and Security ✗ • Human-centered Design and Interoperability ✗

Table 3: Compliance of recent PTSD prediction studies with design principles for AI in general healthcare

Based on the challenges and opportunities highlighted in Section 3.1, the existing principles for AI in general healthcare identified in Section 3.2.1, and the compliance of recent studies on PTSD prognosis prediction with these principles, I underscore the importance of addressing the following key topics when developing design guidelines for AI in mental healthcare:

1. **Stigma and Access to Care**
2. **Enhancing Trust and Evaluate Outcomes**
3. **Understanding Heterogeneity and Complexity of Mental Health Disorders**
4. **Prioritizing Sensitivity to Privacy Concerns**
5. **Advocating Patient-centered Care and Multidisciplinary Approaches**

3.3 Multidisciplinary Collaboration and Human-centered Design Approach in AI Systems Development

The design and development of AI systems, particularly for mental healthcare applications, requires a deep understanding of user needs and the context in which the technology will be applied. This section explores the potential benefits of incorporating multidisciplinary collaboration and HCD approaches in the development of AI systems for mental healthcare, with a specific focus on posttraumatic stress prognosis prediction.

Keywords: Multidisciplinary Collaboration, Interdisciplinary Approach, Human-centered Design, User-centric Design, Participatory Design, Co-design, AI Systems Development, AI in Healthcare, AI in Mental Healthcare

Subject of Study	Articles
Establishing multidisciplinary teams and fostering stakeholder collaboration	• Stokols, Misra, Moser, Hall, & Taylor (2008) • Shortliffe & Sepúlveda (2018) • He et al. (2019) • McDougall (2019) • Kaba & Sooriakumaran (2007) • Shilton (2018) • *The Act* (2021) • Tabassi (2023) • HM Government (2021) • Weiser (2019) • Cherrington et al. (2020) • Baumer (2017) • D. D. Luxton, Anderson, & Anderson (2016)

Continued on next page

Table 4 – continued from previous page

Subject of Study	Articles
Involving end-users early and throughout the design process	• Musen, Middleton, & Greenes (2021) • Yang, Steinfeld, & Zimmerman (2019) • Tulk Jesso, Kelliher, Sanghavi, Martin, & Henrickson Parker (2022) • Cai, Winter, Steiner, Wilcox, & Terry (2019) • Khairat, Marc, Crosby, & Sanousi (2018) • Gundersen & Bærøe (2022) • Cai, Winter, Steiner, Wilcox, & Terry (2019) • Yang, Steinfeld, & Zimmerman (2019) • Tursunbayeva & Renkema (2022) • Gundersen & Bærøe (2022) • Amann et al. (2020) • Shneiderman (2011) • De Silva & Alahakoon (2022) • He et al. (2019) • Simonsen (2004)
Fostering a culture of transparency and trust	• *Ethics guidelines for trustworthy AI | Shaping Europe's digital future* (2019) • Tulk Jesso, Kelliher, Sanghavi, Martin, & Henrickson Parker (2022) • Gundersen & Bærøe (2022) • Barrett, Oborn, Orlikowski, & Yates (2012) • Demerouti (2022) • Mittelstadt, Allo, Taddeo, Wachter, & Floridi (2016)

Continued on next page

Table 4 – continued from previous page

Subject of Study	**Articles**
Committing to continuous evaluation and usability tests	• Tursunbayeva & Renkema (2022) • Singh, Hom, Abramoff, Campbell, & Chiang (2020) • Tang, Lim, Mansfield, McLachlan, & Quan (2018) • Cai, Winter, Steiner, Wilcox, & Terry (2019) • Mohr, Cheung, Schueller, Hendricks Brown, & Duan (2013) • Alzahrani, Gay, Alturki, & AlGhamdi (2021) • Kazdin (2017)

Continued on next page

Table 4 – continued from previous page

Subject of Study	Articles
Addressing political and ethical considerations	• Lee et al. (2021) • Moldoveanu (2019) • Parker & Grote (2022) • He et al. (2019) • Ranchordas (2021) • Weiser (2019) • Mittelstadt & Floridi (2016) • Branch (2019) • *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)* (2016) • Rep. Pallone (2022) • *The Act* (2021) • U.S. Chamber of Commerce (2023) • HM Government (2021) • Harrison, Flood, & Duce (2013)

Table 4: Relevant articles addressing multidisciplinary collaboration and human-centered design approach for AI development in healthcare

3.3.1 Multidisciplinary Collaboration in AI Systems Design and Utilization for Healthcare

Multidisciplinary collaboration is the process of bringing together professionals from different disciplines to work collectively on a project or problem (Stokols, Misra, Moser, Hall, & Taylor, 2008). In the context of AI system development for mental healthcare, this approach can lead to a more comprehensive understanding of the complex and multifaceted nature of mental health, ultimately resulting in more effective and tailored solutions. The numerous articles published addressing the application of AI to healthcare reveal this transition from human-human to human-AI collaboration in the medical field, which requires more than accurate model predictions alone. Indeed, their implementation in practice faces several challenges to ensure clinician adoption of ML-based clinical decision support. Shortliffe & Sepúlveda (2018) outlined these challenges: (1) transparency is required to ensure users' understanding of the tool's recommendations; (2) time is crucial in the busy clinical environment; (3) the use of the tool must be simple, intuitive, and easy to learn; (4) recommendations must be relevant to the clinical context; (5) the delivery of knowledge must respect clinician expertise; (6) recommendations must be based on rigorous science and peer-review processes. Developing further collaboration between designers, clinicians, and patients during the whole design process might respond to most of these challenges.

Indeed, the lack of Human-Computer-Interaction (HCI) considerations is mentioned as a major reason for tool implementation failure (Musen, Middleton, & Greenes, 2021), where clinicians' workflow is not considered enough in the design and implementation of the AI tools (Yang, Steinfeld, & Zimmerman, 2019). Tulk Jesso, Kelliher, Sanghavi, Martin, & Henrickson Parker (2022) demonstrated inconsistent consultation with clinical end users throughout the design process. If users do not discuss and comprehend with designers a tool's capabilities, intended usage, or utility compared to current methods, they may be reluctant to adopt it (Cai, Winter, Steiner, Wilcox, & Terry, 2019). In particular, Clinical Decision Support Systems (CDSSs), which represent one application of AI, migrate with difficulty from labs to clinical practice partly because of user acceptance (Khairat, Marc, Crosby, & Sanousi, 2018). Therefore, the lack of communication with clinicians and the poor integration of their perspectives into the designers' design process seems to be consequent, if not the main, barrier to adoption.

Indeed, multidisciplinary collaboration is first required to ensure that the **objectives** of the designer developing the tool and those of the clinician providing patient care are aligned (He et al., 2019). In the early stages of research and development, clinicians and designers can collaborate to determine what medical tasks might benefit from AI support. The context in which an AI tool intervention is required or at least recommended, based on the patient's conditions, is then made clear to the clinician who has initially advised the intended use of the tool (Gundersen & Bærøe, 2022). Therefore, the clinicians know about the algorithm's objectives, capabilities, and limitations, which leads to better and safer use (Cai, Winter, Steiner, Wilcox, & Terry, 2019). These algorithms' features and subsequent recommendations are relevant to the clinical context and respect clinicians' perspectives.

Moreover, collaborative work can facilitate the tool integration so that the clinicians encounter it naturally in their decision-making **workflow** because they may not know when they need its assistance (Yang, Steinfeld, & Zimmerman, 2019). The clinician can then seamlessly use AI output throughout the decision process. For example, they can integrate it as a second opinion or delegate repetitive tasks to the AI, freeing up more time for responsibilities that require their specialized knowledge and skills (Tursunbayeva & Renkema, 2022). The functionality of the tool, which includes its main steps of usage, is detailed and tailored to align with the clinician's decision-making process (Cai, Winter, Steiner, Wilcox, & Terry, 2019). As a result, it becomes easier for the clinician to learn how to use it. Also, time which is a scarce resource for clinicians is saved which leads to a better chance of tool adoption.

To ensure that the patient is the primary focus and avoid falling back into paternalistic care and 'fixed AI' (Gundersen & Bærøe, 2022) (McDougall, 2019), the patient-centered model can frame the design and implementation processes (Kaba & Sooriakumaran, 2007). **Shared decision-making** can identify the most suitable and effective course of action for patients, using open discussions about its risks and benefits (Amann et al., 2020), and not forcing the patient to blindly trust the clinician or the AI. Collaboration encourages transparency about design and use processes (*Ethics guidelines for trustworthy AI | Shaping Europe's digital future*, 2019), and reinforces mutual trust between the tool, the clinician, and the patient (Tulk Jesso, Kelliher, Sanghavi, Martin, & Henrickson Parker, 2022). Indeed, clinicians must trust the source of evidence supporting their decisions and be able to translate this information in terms of benefits, harms, and trade-offs. Therefore, the interpretation of AI output in clinical practice is determined by including technical and medical point-of-views (Gundersen & Bærøe, 2022), and discussing those with the patient, which makes the tool's recommendations more transparent to its users.

Another advantage of improving communication between these actors is job **feedback**. As AI-based tools provide information and recommendations about diagnoses, prognoses, and treatments, clinicians can use this output to indicate how well they perform their job (Tursunbayeva & Renkema, 2022). On the other hand, clinicians can provide feedback to designers regarding errors in the system (Singh, Hom, Abramoff, Campbell, & Chiang, 2020), interface, and integration to the workflow, and therefore ensure an iterative improvement of the tool (Tang, Lim, Mansfield, McLachlan, & Quan, 2018). Peer-reviewing publications to adjust and validate the technology in a specific context can also provide valuable insight (Cai, Winter, Steiner, Wilcox, & Terry, 2019), and participates in a rigorous scientific process, one in which the design and use of the tool must be grounded.

In addition, policies and ethical guidelines for healthcare services that use AI applications currently lag behind the advancements in AI (Lee et al., 2021). During this prolific development of AI technologies, a widening gap has emerged between politicians' and designers' technical and communication skills (Moldoveanu, 2019). Effective collaboration can lead to better communication of clinicians' and designers' needs and requirements with **policymakers** (Parker & Grote, 2022) and federal agencies such as the Food and Drug Administration (FDA). These latter play a crucial role before tool adoption in refining the system to new technologies, guiding companies through the clinical trial, and establishing processes to foster innovation (He et al., 2019) (Ranchordas, 2021). Also, good communication can help to reach decision-makers in the field of education which, in the mid-term, could add introductory courses to AI and data management to clinicians' curricula. Therefore, it can facilitate their learning of the AI-based tool and reinforce their participation in the design. Lastly, incorporating ethics in the design process of AI systems can help guarantee the responsible application of AI in healthcare settings, adhering to established ethical guidelines and addressing potential ethical concerns proactively (Weiser, 2019).

Finally, implementing a digital breakthrough in a medical system leads to a redefinition of work relations and responsibilities, which may diminish the occupational intellectual and symbolic value of professional groups (Barrett, Oborn, Orlikowski, & Yates, 2012). By promoting collaboration throughout the design process, the **wealth** (including the credit) created by the technology benefits its designers, and users (Demerouti, 2022). It is then clear that AI-based tools respect clinicians' perspectives and assist their decision-making process without replacing them, enhancing transparency and trust associated with the use of AI systems.

In conclusion, the incorporation of multidisciplinary collaboration in the development of AI systems for mental healthcare can offer several benefits, such as:

1. **Comprehensive understanding of user needs**
2. **AI systems tailored to clinicians' workflow**
3. **Facilitated shared decision-making**
4. **Continuous system improvement through feedback loops**
5. **Robust policymaking and adherence to ethical considerations**
6. **Equitable distribution of wealth and recognition of healthcare professionals' contributions**

3.3.2 Aligning Multidisciplinary Collaboration, Human-centered Design, and AI Systems Development

HCD holds considerable importance and relevance in various domains, as it can enhance commercial success, transform education systems, impact daily life, and even influence political stability (Shneiderman, 2011). Designers must strive to enhance AI system performance while considering the ethical implications of their choices, taking into account psychological, social, and political factors that support ethical design practices (Shilton, 2018) (*The Act*, 2021) (Tabassi, 2023) (HM Government, 2021) (Weiser, 2019). When creating AI systems intended for public benefit, incorporating HCD and, specifically, human-centered algorithm design (HCAD) can help prevent discriminatory outcomes. A collaborative and co-design approach is necessary, considering the complex socio-technical systems that align with values, principles, and guidelines for AI system design (Cherrington et al., 2020). Algorithms are deemed superior if their outcomes exhibit enhancement based on predefined performance metrics assessing algorithm accuracy, speed, or computational resources. However, these metrics may deviate significantly from human understanding and interpretations of the intended functions and implications of such algorithms. Such systems must prioritize supporting user agency and fundamental rights to promote a democratic, flourishing, and equitable society. Additionally, allowing for human oversight ensures accountability and ethical use of the technology (Weiser, 2019). HCAD is grounded in three key strategies: 1) theoretical, 2) speculative, and 3) participatory (Baumer, 2017). The theoretical strategy employs behavioral and social science theories for prescriptive and descriptive guidance in algorithm development (feature selection) and interpretation (meaning of algorithm results). The speculative strategy uses imagination to create provocative artifacts that challenge normative values in technology, encouraging exploration and critical thinking. The participatory strategy incorporates users as active participants in the design process, addressing differential expertise and fostering dialectic exchange, leading to more intuitive and user-aligned systems while deepening the understanding of computational imaginaries (the ways people make sense of computational algorithms) (Baumer, 2017).

To effectively integrate multidisciplinary collaboration and HCD into the development of AI systems for mental healthcare, I propose the following strategies. These align with the artificial intelligence life cycle as outlined by De Silva & Alahakoon (2022):

1. **Establishing multidisciplinary teams and fostering stakeholder collaboration**: Assembling teams composed of professionals from various disciplines, such as mental health experts, AI developers, user experience designers, and ethicists, can help ensure that diverse perspectives and expertise are considered throughout the development process (D. D. Luxton, Anderson, & Anderson, 2016) (Stokols, Misra, Moser, Hall, & Taylor, 2008). Encouraging open and transparent communication between the various disciplines is crucial for effective collaboration. Building partnerships among stakeholders, such as healthcare providers, academic institutions, industry, ethicists, and policymakers, can facilitate the sharing of resources, knowledge, and expertise during the AI development cycle (De Silva & Alahakoon, 2022).

2. **Involving end-users early and throughout the design process**: Actively engaging end-users, such as patients, clinicians, and caregivers, in the design and development stages can help ensure that their needs, preferences, and concerns are addressed (He et al., 2019) (Simonsen, 2004). This involvement of users should begin at an early stage, starting from problem identification and data acquisition. It should be maintained throughout the entire AI life cycle, spanning from design and development to deployment and beyond (De Silva & Alahakoon, 2022). Involvement methods may include conducting interviews, focus groups, or surveys to gather input on the system's functionality, interface, and interactions.

3. **Fostering a culture of transparency and trust**: Ensuring that all stakeholders, including end-users and practitioners, understand the underlying principles and decision-making processes of the AI system can help build trust and promote acceptance of the technology (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). Transparent communication about the limitations and potential biases of the AI system is crucial for setting realistic expectations and fostering trust.

4. **Committing to continuous evaluation and usability tests**: Utilizing evidence-based methods and integrating findings from the latest research in mental health and AI can help guide the development process. Employing an iterative development process, with regular feedback loops and continuous evaluations from end-users and other stakeholders, can help refine and optimize the rapidly evolving AI technology to better meet user needs (Mohr, Cheung, Schueller, Hendricks Brown, & Duan, 2013). In addition, performing usability tests with real users can provide valuable insights into how well an AI system meets the needs and expectations of its intended audience and identify potential areas for improvement. Satisfaction, efficiency, and learnability appeared as the most important ones (Alzahrani, Gay, Alturki, & AlGhamdi, 2021). Monitoring and assessing the real-world impact of AI systems on mental healthcare outcomes is essential for identifying areas for improvement. It ensures that the system provides meaningful benefits to users, and helps extend interventions from research to practice (Kazdin, 2017).

5. **Addressing political and ethical considerations**: Incorporating political and ethical considerations from the outset of the development process, through consultations with policymakers, ethicists, and other relevant stakeholders, can help ensure that AI systems are designed with respect for design principles such as user privacy, model explainability, data security, and fairness in mind (Mittelstadt & Floridi, 2016). AI systems must respect current regulations about data (Branch, 2019) (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, 2016) (Rep. Pallone, 2022) and will be required to comply with future AI regulation (*The Act*, 2021) (U.S. Chamber of Commerce, 2023) (HM Government, 2021). Collaboration with policymakers can also help balance regulation and innovation through the creation of experimental spaces for designing AI systems under more flexible rules (Ranchordas, 2021). Designing AI systems with safety and accessibility in mind, considering the diverse needs of users, such as those with cognitive, sensory, or motor impairments, can help ensure that mental healthcare services are inclusive and available to all who need them (Harrison, Flood, & Duce, 2013).

By employing these strategies, AI systems in mental healthcare can be developed with a comprehensive understanding of user needs, while considering ethical implications and ensuring effective communication between various disciplines. This multidisciplinary collaboration and human-centered approach can ultimately lead to AI systems that are more effective, user-friendly, and better suited to addressing the unique challenges faced by individuals seeking mental healthcare.

4 Case Study: PTSD Prognosis Prediction

4.1 Background on PTSD: Prevalence, Diagnosis, Prognosis, and Treatment

The lifetime prevalence of any Potential Traumatic Event (PTE) is up to 80% (de Vries & Olff, 2009) which means that approximately 8 out of 10 people will be exposed to actual or threatened death, serious injury, or sexual violence in different ways during their lifetime (victim, witness, close family member, or friend of a victim). In the hours, days, and weeks following a traumatic event, the survivors may experience psychological reactions including anxiety, sleep disturbance, guilt, emotional numbness, or withdrawal. These can be accompanied by physical reactions such as shakiness, fatigue, poor concentration, palpitations, or loss of interest in sex (Regel & Joseph, 2017). Whether or not these reactions lead to the development of posttraumatic stress disorder (PTSD) depends on a variety of risk factors, including the level of social support, history of previous trauma or mental health problems, relationship issues, and the specific context in which the trauma occurred (Regel & Joseph, 2017). Consequently, in response to the psychological and physical reactions, people tend to follow one of the four trajectories over 12 months post-trauma: (a) resilient class with consistently few PTSD symptoms, (b) recovery with initial distress then gradual remission over time, (c) delayed reaction with worsening symptoms over time, (d) chronic distress with consistently high PTSD levels (Bryant et al., 2015) (I. R. Galatzer-Levy, Huang, & Bonanno, 2018) (Schultebraucks, Sijbrandij, et al., 2021) (Pietrzak et al., 2014) (Bonanno et al., 2012). These commonly observed trajectories are displayed in figure 4.

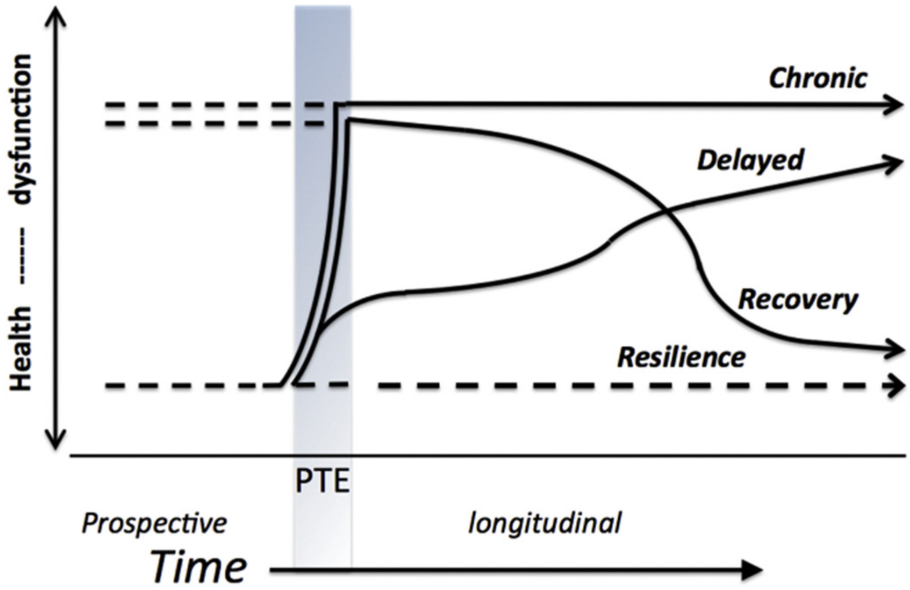

Figure 4: Commonly observed prospective trajectories of response to potential trauma (I. R. Galatzer-Levy, Huang, & Bonanno, 2018)

Most individuals exposed to a PTE do not develop PTSD and recover their sense of equilibrium within a few weeks (resilient class, approximately 65.7%). A significant proportion of those who do develop PTSD initially exhibit high levels of symptoms that remit over time (recovery class, approximately 20.8%). However, a smaller proportion of individuals may exhibit initially low levels of symptoms that increase over time (delayed class, approximately 8.9%), or consistently high levels of symptoms (chronic class, approximately 10.6%) (I. R. Galatzer-Levy, Huang, & Bonanno, 2018) (Lowe et al., 2021). Exposure to PTE can thus lead to the development of PTSD in some cases. PTSD is a common psychiatric disorder, with an average lifetime prevalence of 7.34% across several countries, including Australia, Canada, the Netherlands, New Zealand, and the USA (Dückers, Alisic, & Brewin, 2016). However, it is important to exercise caution when interpreting these prevalence rates, as they can vary widely depending on factors such as study methodology, PTE definition, and population sampling (Knipscheer et al., 2020). Individuals with PTSD can experience severe symptoms such as recurrent distressing memories and/or dreams, dissociative reactions (feeling that the event is recurring), and psychological or physiological distress when exposed to cues that symbolize the traumatic event(s) (Association, 2013). These symptoms can have a significant impact on daily functioning and quality of life. It is important for individuals who experience symptoms of PTSD to seek appropriate treatment and support in order to manage and potentially recover from their symptoms.

PTSD is primarily diagnosed by psychiatric professionals using standardized diagnostic criteria outlined in the diagnostic and statistical manual of mental disorders (DSM-5) in the United States and the international classification of diseases (ICD-11) in Europe. These criteria necessitate the identification of particular symptom clusters, which include experiencing a PTE, intrusive "flashbacks" of the traumatic event, deliberate avoidance of trauma-associated stimuli, negative alterations in cognition and mood, and significant changes in arousal and reactivity. These symptoms must persist for more than one month and cause substantial impairment in social, occupational, or other vital areas of functioning (Association, 2013) (S. D. McDonald & Calhoun, 2010). Guidelines regarding the treatment of PTSD have been identified, ranging from medication to therapy which requires the intervention of a professional therapist, psychologist, or psychiatrist (Watkins, Sprang, & Rothbaum, 2018). However, in addition to the high healthcare costs that accompany these current treatments (von der Warth, Dams, Grochtdreis, & König, 2020), several barriers to mental health treatment exist and must be overcome. Low perceived need for treatment, a desire to handing the problem by oneself, stigmatization, finance and availability, perceived ineffectiveness, and negative experience is among the main reasons why patients do not seek treatment or eventually stop taking it (Andrade et al., 2014). Nearly half of the persons in need of mental healthcare reported no formal healthcare use (Alonso et al., 2007), and people with anxiety-related disorders such as PTSD are very prone to this lack of care (Bijl & Ravelli, 2000). The early study conducted by Hoge et al. (2004) revealed that only 23 to 40% of Iraq/Afghanistan War veterans displaying signs of a mental disorder sought mental health care. In a study conducted among six European countries, the Netherlands showed a higher risk of not using services when there was a need for healthcare (Alonso et al., 2007).

Barriers to the treatment of mental disorders, particularly PTSD, can have serious consequences. These disorders not only significantly impact a person's health-related quality of life (HRQOL) and that of their family and friends, but individuals with PTSD are 80% more likely to experience symptoms of comorbid mental disorders, such as major depressive disorder, substance use disorder, and other anxiety-related disorders, than those without PTSD (I. R. Galatzer-Levy, Nickerson, Litz, & Marmar, 2013). In addition, individuals with PTSD resulting from assaultive violence show a 3-fold increase in the risk of attempted suicide (Wilcox, Storr, & Breslau, 2009). Therefore, it is essential to address the numerous barriers to the treatment of mental disorders, and PTSD in particular, as it represents a significant societal problem. During an interview with Professor Alain Brunet, a researcher at the Douglas Mental Health University Institute affiliated with McGill University, the importance of predicting the prognosis and subsequent trajectories of PTSD as a key feature for finding appropriate treatment was highlighted. Prognosis refers to the prospect of recovering from an injury or disease, as well as a prediction of the course and outcome of a medical condition (Hansebout, Cornacchi, Haines, & Goldsmith, 2009). The effects of PTSD can vary among individuals in terms of the duration and intensity of symptoms (Breslau, 2001), and a better understanding of prognosis and subsequent factors can lead to better decision-making regarding treatment (Hemingway et al., 2013). By improving prognosis and understanding the subsequent factors, healthcare professionals can better tailor treatment to the needs of individuals with PTSD, potentially reducing the barriers to treatment and improving their overall quality of life.

The emergence of AI systems, involving machine learning (ML) and deep learning (DL) techniques, and their ability to detect key features from complex datasets revealed their importance for developing predictive models and thus, could be used to predict prognosis. In recent years, AI development has led to the development of promising tools applicable to the medical field. Such algorithms could significantly improve clinicians' decisions about disease diagnosis, prognosis, and treatment. Techniques such as artificial neural networks (ANNs), support vector machines (SVMs), and decision trees (DTs) have been already applied to the medical field for forecasting the evolution of cancer for instance (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015). Kim et al. (2012) developed a prognostic model based on SVM to predict breast cancer recurrence within five years following surgery. It could help make clinicians' and patients' treatment decisions more consistent and predictable. In particular, DL is a sub-field of ML from which healthcare will also benefit enormously because of the large and heterogeneous datasets generated and the continuous emergence of new medical devices and digital record systems (Esteva et al., 2019). Esteva et al. (2017) used a convolution neural network (CNN) to classify skin cancer from clinical images, which outperforms the average performance of dermatologists. In another study, Shiraishi, Li, Appelbaum, & Doi (2011) used ANN output as a 'second opinion' to assist radiologists' image interpretation. Although AI has found more applications in the fields of oncology, dermatology, and imaging so far (Lee et al., 2021), psychological practice and research will also be significantly impacted by AI technology in mental healthcare. For instance, AI-enabled virtual reality human avatars might help with psychological therapies, evaluations, and testing (D. D. Luxton, 2014). Another research conducted by Shafiei, Lone, Elsayed, Hussein, & Guru (2020) involved participant-based responses to validated questionnaires as the source of input to a CNN to objectively monitor mental health metrics. Hence, AI-based tools can provide robust decision assistance to healthcare professionals in intricate clinical scenarios, including the ones related to mental health. However, limitations to these technologies do exist including biases in real-world data sources and limited predictive power of clinical data alone. AI-based tools, when used in conjunction with human clinicians, can enhance the delivery of care and lead to better outcomes than either could achieve on their own (J. H. Chen & Asch, 2017). Rather than getting caught up in the hype surrounding new technologies, the focus should be on implementing such technologies using design guidelines to ensure their good integration with the mental healthcare system and its stakeholders. Therefore, this paper aims to build design guidelines for implementing AI systems in the mental healthcare context based on the case of PTSD prognosis forecasting.

4.2 Review of the Current Machine and Deep Learning Methods Relevant to PTSD Prediction

ML and DL have gained considerable attention in recent years due to their potential in various applications. This review provides a succinct overview of some of the major ML and DL techniques, namely decision trees, random forests, support vector machines, artificial neural networks, and convolutional neural networks.

4.2.1 Decision Trees

A decision tree (DT) is a fundamental yet potent technique in the realm of ML. It devises a model that takes the form of a tree structure, constituting a binary tree-like diagram. In this model, every node embodies a feature (or attribute), every link (or branch) signifies a decision rule, and every leaf denotes an outcome (Stuart & Peter, 2016). The root node is the prime node in a decision tree, from which outgoing branches traverse to the internal nodes. This traversal follows a Boolean classification until it reaches the leaf nodes. These leaf nodes epitomize all potential outcomes derived from the dataset, as depicted in Figure 5. The allure of decision trees is encapsulated by their simplicity and easy interpretability. However, they also possess certain drawbacks. For instance, when decision trees have numerous extending branches, they can overfit the training data, thereby leading to a lack of generalization capability on unseen data. Moreover, tree stability poses a substantial concern when utilizing tree-based methods. Even minor variations in the data can lead to significant differences in the tree structures (Zhang & Singer, 2010).

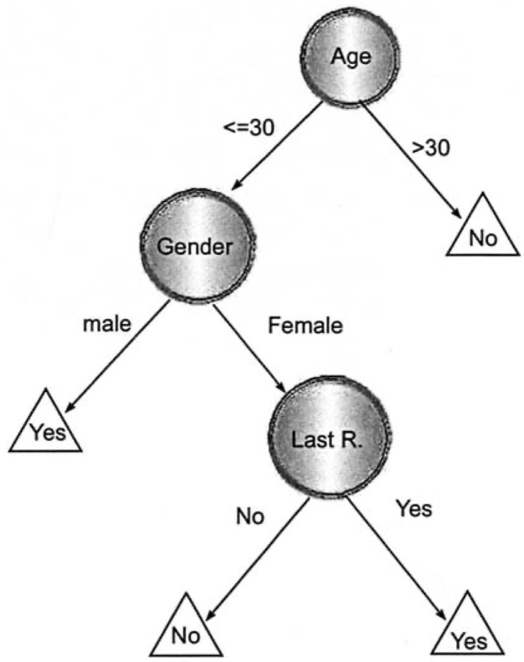

Figure 5: Basic structure of a DT (Rokach & Maimon, 2005)

4.2.2 Random Forests

Random forests (RFs) are an ensemble learning method that incorporates multiple decision trees, thus forming a 'forest'. The main objective of RFs is to mitigate the risk of overfitting and curtail the value of the generalization error (Breiman, 2001). This is achieved by using a technique known as bootstrap aggregation (bagging), which enables the algorithm to repeatedly select random samples from the training set with replacement. It then fits the trees to these samples and averages the predictions across all the trees. Concurrently, RFs employ random feature selection (also known as feature bagging) to diminish correlations between the trees, thereby enhancing the overall predictive accuracy of the model. Due to these combined techniques, RFs improve the model variance, rendering the model more resilient than a solitary decision tree. This is primarily because the averaging process across numerous trees is not susceptible to noise (I. R. Galatzer-Levy, Ruggles, & Chen, 2018). However, despite their benefits, RFs also have a few drawbacks. Their complexity often leads to interpretability challenges, making it difficult to understand the underlying decision-making process. Furthermore, the model demands a substantial amount of computational resources and time for training, which may be a limiting factor in certain scenarios.

4.2.3 Support Vector Machines

Support vector machines (SVMs) are resilient and versatile classification techniques. Their underlying principle involves determining a hyperplane that optimally separates the classes within the input space as shown in figure 6 (I. R. Galatzer-Levy, Ruggles, & Chen, 2018). SVMs are particularly adept at managing high-dimensional data, proving effective, especially in scenarios where the number of dimensions surpasses the number of samples. Despite their strengths, SVMs also have limitations. For instance, they can be inefficient when tasked with training on exceedingly large datasets due to the computational complexity of the underlying optimization problem (Cervantes, Garcia-Lamont, Rodríguez-Mazahua, & Lopez, 2020). Furthermore, SVMs do not inherently offer probability estimates, which can be a drawback in situations where uncertainty measures about the classification decision are desired (X. Wang, Zhang, & Wu, 2019).

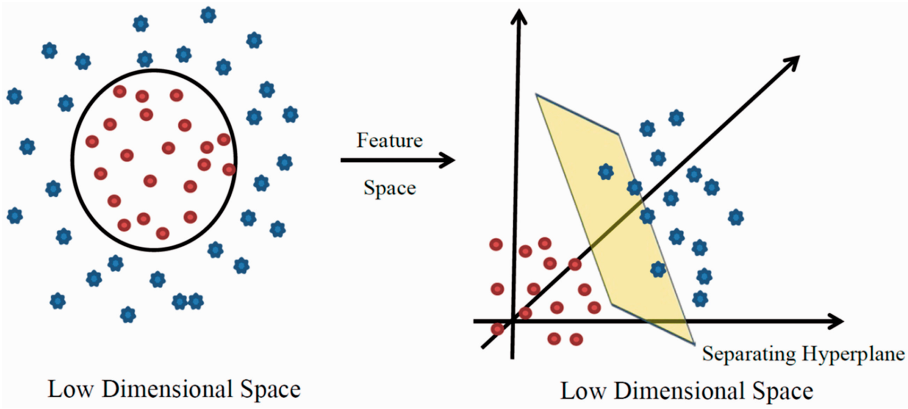

Figure 6: Features that are not linearly separable are mapped into higher dimensional space

4.2.4 Neural Networks

Artificial neural networks (ANNs) serve as a fundamental component of DL, emulating the operational principles of the human brain, as illustrated in Figure 7. Comprised of interconnected layers of nodes or 'neurons', ANNs have the capacity to learn and refine their performance based on experience. In mathematical terms, an artificial neuron is conceptualized as a nonlinear unit that computes a weighted sum of its inputs and subsequently subjects the result to an activation function. The architecture of a feedforward neural network (FNN) encompasses the organization of neurons in a layered structure, with complete connectivity between adjacent layers. The FNN includes an input layer that accepts data, intermediate hidden layers, and an output layer that generates class probabilities in classification tasks (Miotto, Wang, Wang, Jiang, & Dudley, 2018). ANNs possessing more than a single hidden layer are typically referred to as deep neural networks (DNNs). These networks excel at discerning patterns within unstructured data types, such as images, audio, and text. However, their "black box" nature often renders their internal operations challenging to interpret. For instance, DNNs may process input data and yield output without providing transparent explanations of their decision-making processes, such as determining the risk of a mental health disorder. Currently, there exist two primary strategies for enhancing interpretability in such black box models. The first involves altering the input and scrutinizing the resultant changes in the output and neuron activations. The second approach leverages tools to ascertain the contribution of input features to the output, assigning them importance scores (Su, Xu, Pathak, & Wang, 2020). Despite their capabilities, DNNs can be computationally demanding and typically require vast amounts of data to perform optimally. An exemplar subtype of DNNs, convolutional neural networks (CNNs), is specifically designed for image analysis. CNNs employ layers of nonlinear transformations, including convolution-activation, pooling, and fully connected layers, to extract patterns and reduce dimensionality (Su, Xu, Pathak, & Wang, 2020). CNNs have demonstrated exceptional performance in computer vision tasks (LeCun, Bengio, & Hinton, 2015), and show potential for processing clinical images to detect mental health conditions.

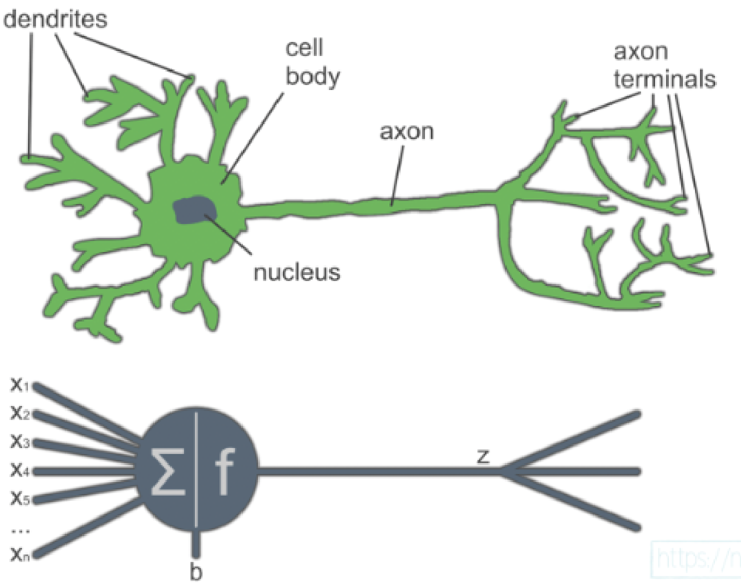

Figure 7: A basic neural network (Kinsley & Kukieła, 2020)

4.3 AI and PTSD Prediction: a Review of Current Tools and Approaches

The article titled "Applications of artificial intelligence - machine learning for detection of stress: a critical overview" by Mentis, Lee, & Roussos (2023) provides a comprehensive evaluation of the implementation of AI systems, specifically ML and DL approaches, in predicting PTSD and stress in a broader context. The study conducted by Wshah, Skalka, & Price (2019) utilized the Gini impurity index to isolate features with high predictive power (Raileanu & Stoffel, 2004). This allowed them to streamline their dataset, focusing on the most relevant early symptoms for predicting PTSD. These symptoms notably included distress related to trauma-related intrusive thoughts, avoidance of trauma-related reminders, negative beliefs about oneself and the world, loss of interest in activities, difficulty concentrating, and sleep difficulties. The study's significant contribution was the prediction of elevated PTSD levels one month after a traumatic event. By employing self-reported symptoms and ML models like logistic regression (LR), naive bayes (NB), SVMs, and RFs, the authors demonstrated that automated predictions could assist healthcare providers in identifying the need for early interventions 10 to 20 days post-trauma. The ML model achieved an area under the curve (AUC) value of 0.85, indicating its effectiveness in predicting substantial PTSD symptom escalation based on observable symptoms. These findings underscore the broad utility of symptom tracking via smartphone data, suggesting that smartphone surveys for self-reported symptoms can be further streamlined. Despite these promising results, the study acknowledged limitations that include the need for implementation in real clinical settings and the potential benefits of incorporating additional data such as baseline PTSD symptoms, demographic variables, and trauma histories.

Similarly, (Saxe, Ma, Ren, & Aliferis, 2017) employed SVM, RF, and Lasso techniques to predict PTSD using a dataset comprising 163 hospitalized children aged 7-18 who had experienced an injury. From the 105 risk factor variables initially considered, ten variables consistently surfaced as influential factors across multiple bootstrapped samples, reinforcing their reliability and stability. These variables included prior history of PTSD, prior externalizing symptoms, prior experiences of loss, acute stress and pain, protective factors, candidate genes, prior help-seeking, ketamine administration, and clinical symptoms. Despite achieving a mean AUC value of 0.79 for predicting PTSD, the authors noted study limitations such as the relatively small sample size, the specific focus on injured children, and the narrow time frame for PTSD prediction.

Further studies have applied SVM techniques to predict PTSD using data collected from trauma survivors shortly after their admission to the emergency department (ED) (Karstoft et al., 2015) (I. R. Galatzer-Levy, Karstoft, Statnikov, & Shalev, 2014). (Schultebraucks et al., 2020) outlined the development and validation of an algorithm designed to predict the 12-month PTSD trajectory. The algorithm utilized data that is typically collected from electronic medical records and is supplemented by brief clinical assessments of the patient's immediate stress reaction. The results illustrated the algorithm's externally validated accuracy in discriminating PTSD risk with high precision. They mentioned that future research should aim to broaden its applicability to the diverse and clinically heterogeneous ED population in the context of routine medical care. These studies contribute to the identification of unique predictors that impact the trajectory of PTSD, such as cortisol response as a long-term predictor of distress and resilience (I. R. Galatzer-Levy, Steenkamp, et al., 2014). I. R. Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev (2017), using SVM and feature selection techniques, identified sets of features that incorporated clinical, neuroendocrine, psychophysiological, and demographic information. These features were crucial in creating robust classifiers that accurately determined membership in specific trajectories, achieving an AUC value of 0.82. The study also revealed a unique pathway linking childhood trauma to non-remitting PTSD, mediated by lower cortisol levels during ED admission. These findings add to our understanding of PTSD's multifaceted nature, illuminating the variety of pathways individuals may follow when exposed to stress and danger (I. Galatzer-Levy, Bonanno, Bush, & LeDoux, 2013). The application of ML techniques and data science facilitates the development of methods for identifying risk factors, classifying individuals' health status, comprehending mechanisms of change, and formulating effective treatments (I. R. Galatzer-Levy, Ruggles, & Chen, 2018).

In a related study by (Schultebraucks, Qian, et al., 2021), the emergence of PTSD was anticipated based on data assembled before the deployment of military personnel to Afghanistan. ML methodologies, particularly RF and SVM algorithms, were utilized to estimate the probability of a provisional PTSD diagnosis within 90-180 days following a 10-month tour of duty. Both the RF and SVM models registered AUC values of 0.78 and 0.88, respectively, which affirmed their effectiveness in precisely predicting PTSD outcomes. Furthermore, these ML techniques were also used to discern the longitudinal trajectories of PTSD symptoms, where the RF and SVM models achieved AUC values of 0.85 and 0.87, respectively. Influential predictive attributes that prominently figured in these models encompassed pre-deployment sleep quality, anxiety, depression, sustained attention, and cognitive flexibility. The integration of blood-based biomarkers such as metabolites, epigenomic markers, immune markers, inflammatory markers, and liver function markers, offered additional insights that augmented the most powerful predictors.

In the same spirit, A. D. McDonald, Sasangohar, Jatav, & Rao (2019) conducted a study aimed at investigating the feasibility of monitoring active-duty military personnel, a demographic at increased risk of developing PTSD, between sessions. The researchers concentrated on analyzing real-time heart rate data from 107 veterans, procured through wearable sensors, with an eye toward early detection of PTSD triggers. This study employed an array of ML algorithms including DT, SVM, and RF, alongside DL models such as Neural Networks and CNN. This research bridged a crucial gap in the field by leveraging non-intrusive sensing, user-centric, and persuasive design principles, coupled with ML and DL methodologies, to engineer a perpetually active intervention to support veterans suffering from PTSD. The findings disclosed that the SVM, RF, and CNN algorithms displayed superior predictive capabilities in terms of PTSD symptom onset when compared to random classifiers. The corresponding receiver operating characteristic curve (ROC) curves and AUC values are depicted in figure 8. Analysis of these predictions unveiled a positive correlation between elevated heart rates and the emergence of PTSD triggers as delineated by the implemented algorithms.

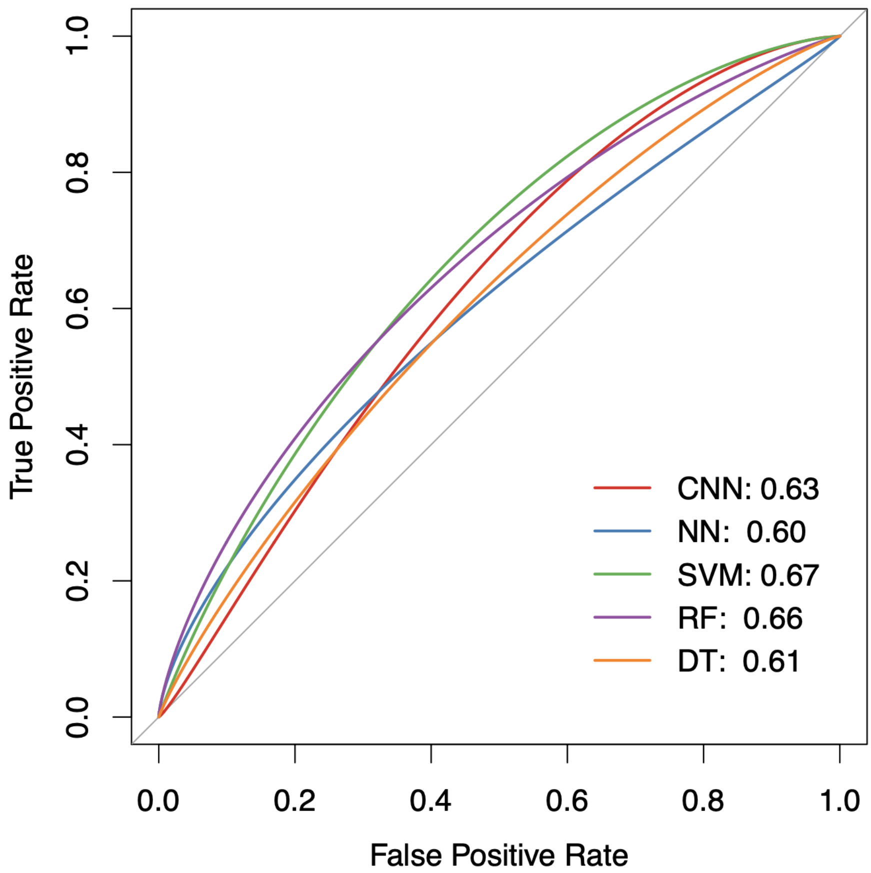

Figure 8: ROC curves and corresponding AUC values for the 5 algorithms in the study conducted by A. D. McDonald, Sasangohar, Jatav, & Rao (2019)

Although burgeoning research underscores the potential of AI, encompassing ML and DL techniques, in the prediction and comprehension of PTSD, these studies concurrently emphasize the necessity for larger and more diverse samples, as well as the incorporation of various data types to enhance prediction accuracy. Future investigations might benefit from examining the potential of these technologies in real-world clinical scenarios, thereby further substantiating their applicability and efficacy in early intervention strategies. Moreover, the deployment of AI systems into healthcare could facilitate a more personalized approach, enabling the identification of individuals at risk, and ensuring that they receive appropriate and timely care. The continual advancement of our understanding of PTSD necessitates a corresponding progression in the development of tools and the utilization of techniques to design them. Consequently, it becomes imperative to design and deploy AI systems within the field of mental healthcare to enhance our ability to predict, comprehend, and effectively manage this intricate and multifaceted disorder.

5 Development of Design Guidelines for integrating AI into Mental Healthcare

5.1 AI System Design Cycle

The article published by De Silva & Alahakoon (2022) depicted a complete AI life cycle consisting of three main phases subdivided into several steps as shown in figure 9: (1) design, (2) develop, (3) deploy. Different human expertise is required at each phase of the overall design process.

Figure 9: The CDAC AI life cycle: Three phases of (1) design, (2) develop, and (3) deploy and 19 stages (De Silva & Alahakoon, 2022)

Inspired by this cycle, the following key steps are considered to build and organize the design guidelines for integrating AI into mental healthcare:

1. **Problem Definition and Stakeholder Engagement**: Define the problem the AI system will solve in the context of mental health, and engage stakeholders to ensure the system meets their needs.

2. **Acquisition of Data and Its Challenges**: Identify and collect the necessary data for model training, and prepare it to be used effectively.

3. **Ethical and Legal Considerations in Data Collection and Use**: Address potential biases in the data or model and ensure the privacy and security of user data. Design the interface and interactions for the AI system, ensuring it complies with users' requirements and relevant regulations.

4. **Design, Selection, and Interpretability of Models**: Choose the most appropriate AI model or approach for the problem at hand.

5. **Model Evaluation, Risks, and Impact**: Train the selected model using the prepared data, and evaluate its performance.

6. **System Deployment, Use, and Maintenance**: Deploy the AI system, ensuring plans are in place for regular maintenance and updates.

7. **Feedback and Improvement**: Incorporate mechanisms for feedback, learning from the AI's performance, and improving the system over time.

5.2 Interviews Analysis

This subsection presents the analysis of the interviews I conducted for this study. The participant pool includes a diverse range of professionals — clinicians, researchers, engineers, and organizational directors — from a total of 7 different countries. The following table 5 indicates the profession and the country of work related to each participant.

Participant	Profession	Country of work
1	Psychiatrist	Germany
2	Clinical psychologist	United States
3	Biostatistician	France
4	PTSD researcher	Canada
5	Research scientist (tech company)	United States
6	Director (healthcare Sector)	Ireland
7	Mental health researcher	Canada
8	Clinical director (mental health sector)	Canada
9	Director (traumatic stress studies)	Australia
10	Cognitive AI researcher	Netherlands

Table 5: Profession and country of work of each participant

The interviews were conducted using a semi-structured interview guide, which is provided in appendix B.3. This guide was meticulously organized to align with the AI system design cycle, illustrated in subsection 5.1. This alignment ensures comprehensive coverage of the entire AI system design process, from the initial problem definition to feedback and improvement.

The procedure and interview guide were reviewed and approved by the ethical committee of TU Delft, ensuring adherence to all ethical guidelines. Each participant was provided with a 'Participant Information' document detailing the purpose and process of the research. Participants were made aware that their involvement was voluntary, and that they had the right to withdraw at any time. Following this, all participants gave their informed consent by signing an 'Informed Consent Points' form. Both documents can be found in appendix B.4.

The information gleaned from these interviews is presented and examined in this section, shedding light on the diverse perspectives and insights these professionals have provided toward my research question.

1. **Defining the Problem and Engaging Stakeholders**: The complex nature of mental health care, highlighted by a clinician's statement that, "Mental health is far less defined than diabetes or cancer; there is a high degree of subjectivity", underscores the nuanced problems faced in this field. An array of mental health disorders exist, including depression, anxiety, PTSD, and psychoses. These disorders are categorized based on a diverse array of symptoms and their varying severity levels, as outlined in the DSM. "Depression, anxiety, PTSD, and psychoses are defined by the DSM and people can be labeled or not according to symptoms and their severity", one clinician expounds. Another important aspect of mental healthcare is the role of individual circumstances. As one clinician explained, "Everybody gets trauma in their life, but some people suffer from it. Most people with a car accident, or child sexual abuse don't have mental health disorders, but they suffer". This was expanded upon with the assertion that socio-economic factors considerably influence mental health outcomes. "Certain individuals do worse overall because they don't have social support

systems, or because they have so many life stressors, because of their economic situation, their interpersonal situation".

With an acknowledged need for assistance in the mental health sector due to the system's current strain as mentioned by a clinician, "Overall, there is a good awareness within the mental health sector that they need some help. The system is currently overloaded", it's vital to engage with stakeholders such as clinicians, patients, and engineers. One clinician emphasized, "Involving clinicians would help them understand what the system is basing its decisions on, and it leads to some extent to higher trust but mostly better trust". This participatory design approach, despite its challenges, is highly recommended. This same clinician also pointed out the potential pitfalls, "Participatory design requires that you spend time with people in designing, redesigning, and co-designing... As much as people would like to, they don't necessarily have the time to engage in that process... It can be very tricky to do but it's certainly best practice".

While focus groups have proven effective in this process, their composition often depends on the project specifics. A clinician stated, "We started with focus groups with clinicians... At the end of the day the patient can use the system, but they were not included in the design of it for a lot of practical and ethical reasons". However, the inclusion of patients, especially those who have recovered, can provide valuable insights into the user experience. "It might be good to talk to people who had already gone through it... At the end of the day, these are the people who need the things, and some don't have a computer and don't know how to install the software. You can have low computer literacy, so they should be included in the design."

2. **Acquisition of Data and Its Challenges**: Integral to AI application in mental healthcare is effective data acquisition. Clinicians' expertise plays a significant role, as one explained, "There are some things that clinically you can detect that maybe you may not see in the data". Despite this, acquiring adequate quality and quantity of data suitable for AI applications remains a significant hurdle, with a clinician expressing, "Quality and quantity of data is a big challenge for deep learning. Healthcare data are collected through validated measures but with very different constructs". Routine treatment data and medical records offer a potential wealth of information for predictive pattern analysis. However, this data collection often falls short, as another clinician notes, "One of the benefits of using medical records is that medical labs and emergency rooms are fairly standard, but we had a lot of work we had to do because for any given subject you don't have complete data". This calls for a more systematic approach to data collection in mental health, highlighting the necessity of bespoke data collection as "Most data out there is not structured to answer this kind of question".

   The clinician voices concern over the applicability of deep learning techniques to healthcare data due to the challenges in quality and quantity of data, especially when it comes to mental health disorders, "Humans are so complex, what if there is something more that is happening to the patient." Therefore, clinician input is vital at this stage, as they possess the capability to discern nuances and factors that may not be evident in the data, but are critical to understanding patient health. Although comprehensive data collection is necessary for recognizing patterns in patient visits and capturing complete treatment courses, it is often impeded by diverse data sources and health insurance company limitations. Hence, an increased emphasis is needed on data accessibility and integration.

3. **Ethical and Legal Considerations in Data Collection and Use**: Understanding the ethical and legal parameters of data collection and usage in AI applications for mental health is essential. One clinician highlights the challenge of data accessibility: "Information is in a lot of different places right now unless it is something like the Department of

insurance affairs which is hard to get in". Data privacy is a critical issue for patients, as stated by a clinician, "A lot of patients we work with would be concerned by data privacy; our population is very untrusting". Ethical dilemmas also emerge around the potential for clinician practice change, which could result in misconduct, given access to this data: "If a person is not at the highest risk, it brings a lot of ethical concerns. Clinician practice is going to change and if the clinicians know this information, they can be responsible and they can misconduct".

AI tools must effectively maneuver ethical conundrums and mitigate bias, especially when their prognoses may challenge the clinical judgment. Furthermore, transparency in the treatment selection process should be mandatory, and patient discussions should emphasize shared decision-making and trust-building. "Consent, particularly around data privacy, is vital". Interestingly, tools such as chatbots may circumnavigate some ethical and legal concerns as “they are not medical tools, people have not been diagnosed yet.”

4. **Design, Selection, and Interpretability of Models**: Designing an AI model for mental health encompasses several key considerations. One clinician identified a challenge: "One challenge with Machine Learning solutions is not to dismiss what seemingly are uninterpretable patterns". Moreover, it’s crucial not to unduly limit Machine Learning’s potential by conforming rigidly to the views of clinicians and researchers: "It’s important not to get too caught by the particular views of clinicians and researchers because it reduces the potential of Machine Learning". AI can play a significant role in mental healthcare, as long as we recognize uninterpretable patterns and remain aware of inherent biases: "Clinical decision-making is a very interesting and complex process both in terms of human cognition, understanding systems, and dealing with biases in clinical practice."

   It’s necessary to engender an appropriate level of trust in the system, as one clinician suggests, "You want an appropriate level of trust in the system, if they don’t understand how it works, it could lead to either clinician saying that they must have overlooked something, and the system must be right or that stupid things come out from the system and they must ignore it". Mitigating false negatives should be prioritized as "the idea of missing anyone is not acceptable for clinicians." Transparency about the system’s capabilities and limitations is vital: “The challenge is to be explicit about what you can and cannot do... In some cases, such systems are trained to sound good, not to be right.” Ensuring the system is user-friendly and comprehensible to those without statistical training is paramount, "You must be sure that people are still realizing that they are talking to a system”. The system’s success hinges on its usability: "At the end of the day, the full success of a system is defined by how well people can use it to make better decisions, and that depends both on how well the system is operating and how the people use it". Thus, balancing system accuracy with usability is crucial. Emphasizing clinical judgment and providing interpretable results can help manage expectations. As one clinician noted, "For patients, rather than only accurate rates, it’s important to give the meaning and the options according to the results."

5. **Model Evaluation, Risks, and Impact**: Evaluating AI models in mental health involves reconciling enthusiasm for the potential of the technology with an understanding of its limitations. A clinician shares, "I think it’s exciting, but my concern always is if something is not 100%, it can be used to guide, to inform but still using clinical judgment". Concerns about missed diagnoses resonate, with a clinician noting, "The idea of missing anyone is not acceptable for clinicians. Even if you capture people who are wrong-diagnosed and false positive, it’s better than missing people". Ensuring patient safety is of the utmost importance: "You want to guarantee the patient’s safety. Embedding the system in such a way that all the impacts it has, are used in a safe way would be important".

Recognizing the limitations of AI tools and communicating them transparently to patients and clinicians is critical to mitigating the risks of misuse. Moreover, how model outcomes are communicated to users is pivotal in mitigating risks associated with patients' reactions to these outcomes: "The phrasing matters a lot... If the tool is used on the patient's side, it can advise the people by saying: based on what we know about people similar to you, we see that often this happens. Therefore, it might be different for you, but we might advise you to talk to someone."

6. **System Deployment, Use, and Maintenance**: The deployment of AI systems in mental health care can encounter various challenges, such as a lack of enthusiasm from hospitals for validation and implementation in real-world settings. Transitioning from development to practical application can be a daunting task: "Nobody is really interested in integrating it because there is not much incentive to know who is going to develop PTSD. The risks are very long-term. With PTSD, the costs are spread out over a long time. There is not a clear benefit to anybody, except for the patient of course."

   During deployment, several factors need consideration. Ensuring patient safety, clinician understanding of the system's limitations, and careful crafting of system advice to patients are essential. Additionally, timing and the environment of deployment are key. Creating a customer base for the system and offering clear benefits for all users, not just patients, is also vital. One clinician emphasized the lack of incentive in some medical contexts: "In the context of the emergency room, now we can tell at a high probability who will have PTSD... But there was not a lot of interest because you first have to think about the customer experience. Who wants that tool and what benefits you can get out of it? In the context of medicine, it has a lot to do with who is responsible for treatment."

   For successful deployment, especially in lower-income countries, cultural and language adaptation is crucial: "There is a need for this kind of cultural variation in people's understanding of psychological distress, and how they express psychological distress. There is a lot of mental health stigma in Europe but it's a lot more in certain parts of the world, in the Middle East you don't talk about suicide."

   A substantial number of clinicians recognize the benefits of such technology: "There is always a couple of people who don't believe that something that has no human in the loop could ever be effective or help people, but it's outweighed by the people who say they need something." AI interventions can help more people, including those uncomfortable with direct human interaction, on waiting lists, or hesitant to seek a clinician: "This could be a first step for people on the waiting list or people who don't want to talk to humans... A chatbot is anonymous and you can pull out at any time you want." AI systems can also assist clinicians by providing a second opinion: "Clinicians are not very good at predicting outcomes for people, they are very poor actually... It can be used as a second opinion."

   Larger-scale applications can be beneficial too: "It might be also helpful for general practitioners, family doctors, to remind them to ask about specific stuff... It can also be used with research organizations because, over time, massive data collection will allow us to see if it needs changes in the way we diagnose the symptoms."

7. **Feedback and Improvement**: AI systems in mental healthcare are dynamic tools that require continuous updating, improvement, and evaluation based on feedback from clinicians and patients: "There are a lot of practical challenges in applications in the medical domain which has to do with regulation, law, and timelines so that's one of the other things a system like this should always be evolving." Clinician input is essential to the iterative evolution of these systems. This phase also involves sharing successful case studies to build further trust and demonstrate the effectiveness of the tool.

5.3 Innovative Design Guidelines for AI Systems

In my endeavor to develop AI systems tailored to address the complex domain of mental health issues, spanning the entire continuum from diagnosis, passing by prognosis, to therapy management, I have diligently crafted these guidelines listed in table 6. These have been compiled based on extensive literature reviews that I conducted, through which I discerned the manifold challenges and opportunities intrinsic to the employment of AI in mental healthcare solutions. In this process, I scrutinized the extant design principles, guidelines, and best practices within the larger healthcare context. I also integrated insights derived from collaborative, multidisciplinary approaches and human-centered design principles, underscoring my belief in the centrality of the user in creating effective solutions. I enhanced the depth and applicability of these guidelines by assimilating findings from a case study on PTSD and insights gleaned from interviews with clinicians and designers actively participating in the mental healthcare domain. I have designed these guidelines to align with the lifecycle of the AI system, beginning from problem definition through to feedback and iterative improvements.

However, it is pivotal to acknowledge that mental healthcare is a diverse and multifaceted field. Consequently, while these guidelines serve as a robust framework, they cannot provide a universally applicable blueprint. The specific context and requirements of each system necessitate distinct adjustments and considerations, further complicated by the inherent complexities of the mental healthcare context. This gives rise to certain limitations, addressed in subsection 5.4. Hence, these guidelines should be applied judiciously, with adaptations made to correspond to the distinct requirements and circumstances of each project

Design cycle phase	Guidelines
Problem Definition and Stakeholder Engagement	1. Define a precise and measurable problem statement focused on a specific aspect of mental health. The problem statement should serve as a direct call to action. 2. Engage with a diverse group of stakeholders, including mental health professionals, patients, caregivers, and policy-makers, to ensure the solution addresses actual needs and incorporates multiple perspectives.
Acquisition of Data and its Challenge	1. Identify and obtain a diverse set of data variables relevant to the mental health issue, including demographic information, clinical history, and psychosocial context. Use multiple data sources where feasible. 2. Address data limitations proactively. Build strategies to handle missing or biased data and ensure data quality and representativeness.

Continued on next page

Table 6 – continued from previous page

Design cycle phase	Guidelines
Ethical and Legal Considerations in Data Collection and Use	1. Design robust protocols for data privacy and security, incorporating transparency measures so users are aware of how their data is used and protected. 2. Actively seek and mitigate potential biases in both data and models. Use methodologies to regularly audit system fairness.
Design, Selection, and Interpretability of Models	1. Select an AI model prioritizing transparency and interpretability. Users should be able to comprehend the model's logic, thus instilling trust in the system. 2. Ensure the system interface is intuitive, user-friendly, and accessible, providing a positive user experience.
Model Evaluation, Risks, and Impact	1. Use a diverse set of evaluation metrics, encompassing statistical performance and qualitative impacts, such as user satisfaction and real-world effectiveness. 2. Establish risk assessment mechanisms, preparing for various scenarios including false positives and negatives. Create strategies to mitigate identified risks.
System Deployment, Use, and Maintenance	1. Create a comprehensive deployment plan considering the unique contexts and workflows where the system will be used. 2. Develop a routine maintenance protocol that checks for data drift, system performance, user satisfaction, and regulatory changes.
Feedback and Improvement	1. Establish continuous channels for user feedback and monitor user interaction with the system to identify potential improvements. 2. Adopt a culture of iterative development, continuously refining the system based on quantitative performance metrics and qualitative user feedback.

Table 6: Design Guidelines for Integrating AI Systems into Mental Healthcare

5.4 Limitations to the Applicability of the Guidelines

Concerning the guidelines proposed, while they offer a comprehensive roadmap for integrating AI into mental healthcare, they have inherent limitations. A significant challenge lies in achieving meaningful stakeholder participation. Despite the critical importance of diverse input, engaging stakeholders for design input, feedback, or consultation can be challenging due to constraints on time, resources, and the sensitive nature of mental health issues.

The acquisition of a comprehensive and representative dataset also presents hurdles. Mental health data is highly complex and obtaining data that reflect all socio-economic subpopulations is difficult. These challenges arise from issues such as varying access to healthcare services, diverse health-seeking behaviors, and different levels of comfort with technology, which can create disparities in the data collected.

Navigating the dynamic landscape of data privacy and AI regulation is another complex aspect. With new laws and protocols emerging, developers need to stay vigilant and responsive, adapting these guidelines as necessary to ensure compliance.

Model selection and interface design are also nuanced areas. Choosing interpretable AI models can limit the predictive power of the system. Moreover, designing user-friendly interfaces that cater to a spectrum of users, from clinicians to patients with varying levels of digital literacy, poses a unique challenge.

In the context of model evaluation, a significant limitation lies in the existing metrics' inability to thoroughly assess an AI model's effectiveness in practical settings. Although we acknowledge the need for more comprehensive metrics, creating these measures presents challenges. These should capture multiple dimensions, including patient satisfaction, ease-of-use, ethical and legal considerations, and the impact of false predictions. However, the task of quantifying such aspects can be complex and subjective, making it a pressing yet challenging necessity..

Finally, maintaining a system that is responsive to continuous monitoring, user feedback, and iterative development is resource-intensive. The ever-changing nature of mental health disorders demands consistent adaptation of the system to maintain performance.

In sum, these guidelines, while valuable, should be adapted to the specific context of individual projects. They should be viewed as a dynamic framework, requiring ongoing updates in response to the evolving legal, ethical, and technological standards.

6 A Virtual Agent for Post-traumatic Stress Early Detection and Support

6.1 Introduction to Virtual Agents and Applications

Virtual agents or chatbots have emerged as promising tools in healthcare, offering support and advice. Whether integrated into virtual reality environments or as pop-ups on computer screens, these agents have the potential to provide a virtual 'coach' experience. Online users can access these agents to receive answers to their clinical queries, discuss treatment options, find suitable treatment centers, complete self-assessment questionnaires, and enjoy a human-like interaction, thus enhancing healthcare awareness and accessibility (Rizzo et al., 2016). In the context of mental health disorders and post-traumatic stress, these virtual agents or chatbots hold significant promise. They can contribute to early detection of symptoms, make preliminary predictions about prognosis, and identify individuals at risk. The early study conducted by Hoge et al. (2004) revealed that only 23 to 40% of Iraq/Afghanistan War veterans displaying signs of a mental disorder sought mental health care. This hesitancy, particularly among those who need it most, highlights a critical area requiring attention in military mental health care. Furthermore, effectively disseminating healthcare information to military service members, veterans, and their loved ones remains an ongoing and escalating challenge. Despite the increasing availability of medical information online, many individuals find the process overwhelming, contradictory, and impersonal. The primary obstacles to treatment include the limited availability of mental health specialists, stigmatization of mental health services, and difficulties in accessing care due to long waiting lists, restricted clinic hours, inadequate referral processes, and geographical constraints (Rizzo et al., 2016). In response to these challenges, the design of SimCoach allows users to engage in interactive dialogues about their healthcare concerns with an intelligent virtual agent as shown in figure 10.

Figure 10: SimCoach, a virtual agent healthcare guide (Rizzo et al., 2016)

These virtual characters use speech, gesture, and emotion to introduce system capabilities, gather anonymous information, provide support and advice, offer relevant online content, and potentially guide users toward seeking appropriate care. The implicit goal of the SimCoach project is to help users recognize their need for care and initiate the process of psychological or medical treatment. The aim is also to allow SimCoach to access databases including patients' electronic medical records (EHRs). SimSensei is another example of a virtual human platform able to interpret real-time audiovisual behavioral signals from users interacting with the system. Both of these applications do not aim to substitute a human presence in clinical activities but rather serve as a solution when the availability of a real person is limited. The capabilities of such virtual agents encompass the following (Hudlicka, 2016):

1. **Improve the distribution of evidence-based treatment**
2. **increase treatment accessibility**
3. **Aid in treatment continuity between sessions**
4. **Adapt to individual requirements and cultural preferences**
5. **Promote engagement while enhancing motivation**

These findings extend beyond the military sector as nearly half of individuals in need of mental healthcare, as reported by Alonso et al. (2007), did not access formal healthcare services. This lack of care is especially prominent among those with anxiety-related disorders such as PTSD (Bijl & Ravelli, 2000).

In healthcare settings, the integration of virtual agents directly into clinics also holds promise. Hypothetically, a virtual agent could be deployed at a clinic, awaiting your arrival and conducting an initial clinical interview. By gathering vital information, the virtual agent could initiate the treatment planning process. Utilizing advanced sensing technology, such as cameras and microphones, the virtual agent could 'observe' your facial expressions, body language, and vocal cues. This data could potentially assist in inferring your psychological state, thereby enhancing the interaction and providing valuable documentation of your mental status over time (Rizzo et al., 2016). It is important to note that the aim is not to replace well-trained clinicians but rather to assist them in their clinical decision-making process. By gradually exposing phobic individuals to their fears through simulations, virtual reality has proven effective in assisting patients in overcoming specific phobias, including public-speaking anxiety (Anderson, Zimand, Hodges, & Rothbaum, 2005).

Therefore, the field of clinical therapy stands on the precipice of an exciting transformation as it begins to incorporate autonomous virtual human agents into practice. In the past decade, significant strides have been made regarding such technologies and their pragmatic applications. As computational power increases and progress in artificial intelligence, graphic rendering, animation, speech recognition, and natural language processing continues, it is anticipated that the role of virtual agents in clinical settings will become increasingly significant. One such innovative tool, specifically conceived for this thesis, is AnchorAid, which offers virtual support in post-trauma recovery. The following discussion will introduce and explore AnchorAid, offering insights into how this tool can enhance trauma support outcomes.

6.2 AnchorAid: My Hypothetical Virtual Support Framework for Post-Trauma Recovery

The Rationale behind AnchorAid: AnchorAid represents an innovative tool, theoretically proposed within this thesis, that endeavors to address the challenges and harness the opportunities associated with the integration of AI in mental healthcare. In this section, I delve into the theoretical design of AnchorAid, demonstrating its potential use and the applicability of the guidelines proposed in subsection 5.3. This tool, in the form of an accessible online virtual assistant, strives to break down barriers such as stigmatization, prolonged waiting lists, and geographical constraints that often dissuade individuals from seeking professional help following a traumatic experience. Again, AnchorAid does not aim to replace well-trained clinicians. Instead, it seeks to provide early personal support to users and assist clinicians along the patient management process, as shown in figure 11. Through its interactions with users, AnchorAid collects data that can equip clinicians with valuable insights into a patient's overall health status and history.

By utilizing an evolving virtual chatbot interface — with the potential to advance into more sophisticated forms such as 3D-designed virtual agents or virtual reality — this service provides personalized feedback and support concerning potential post-traumatic stress symptoms. Notably, AnchorAid leverages natural language processing capabilities to analyze an individual's narrative, a methodology currently under examination for its potential in predicting depressive episodes. Further, by accessing and processing diverse data sources, such as just mentioned, natural language, but also EHRs, and wearable sensor data, AnchorAid encapsulates the complexity of symptoms. In doing so, it can unearth hidden predictors of adverse prognosis concerning post-traumatic stress trajectories. AnchorAid is thus designed to offer assistance in the early stages of post-trauma recovery. It addresses clinical queries, provides initial symptom management recommendations, helps users complete self-assessment questionnaires, and identifies those at risk, thereby guiding them toward appropriate treatment centers. Through the seamless transmission of information collected during user interactions with the virtual assistant, AnchorAid enhances and streamlines the decision-making process for clinicians, should the user decide to seek professional help in response to the recommendations provided by the tool. Ultimately, this tool aspires to augment the accessibility of mental healthcare services worldwide, particularly in low- and middle-income countries.

Nevertheless, I acknowledge that the design of AnchorAid must address substantial risks associated with data privacy, regulatory compliance, patient safety, acceptability, and usability. These concerns necessitate a rigorous approach to the design of AnchorAid, incorporating strict guidelines. The objective of the following table 7 is to envision how a team of designers, including myself, could respond to these guidelines in the specific case of designing AnchorAid.

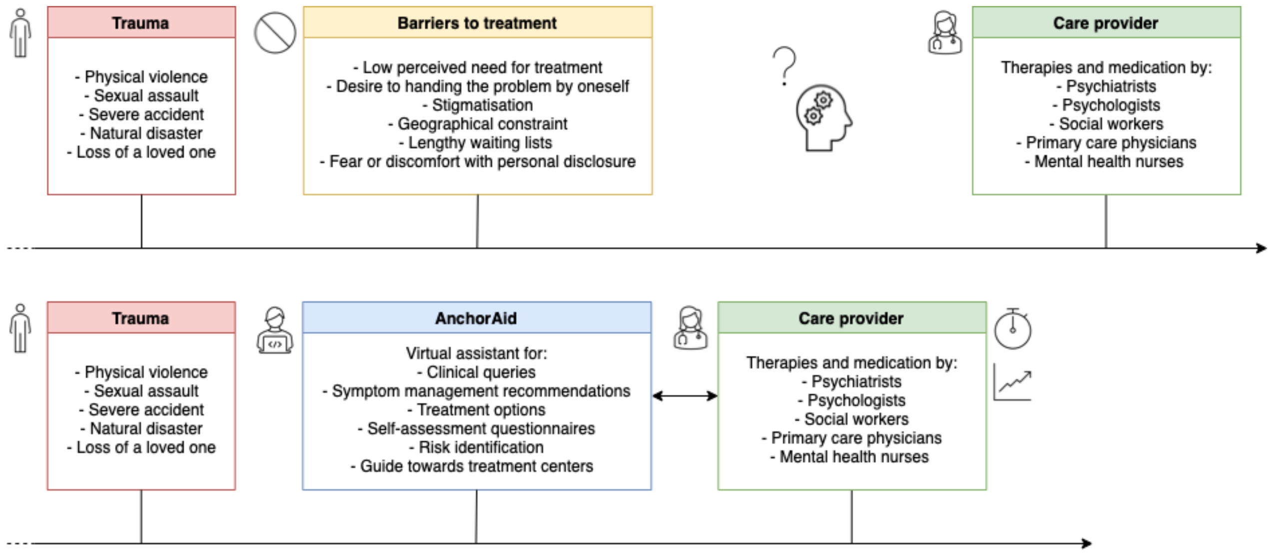

Figure 11: AnchorAid, a virtual agent integrated into patient management workflow

Design cycle phase	Guidelines	AnchorAid Guidelines
Problem Definition and Stakeholder Engagement	1. Define a precise and measurable problem statement focused on a specific aspect of mental health. The problem statement should serve as a direct call to action. 2. Engage with a diverse group of stakeholders, including mental health professionals, patients, caregivers, and policy-makers, to ensure the solution addresses actual needs and incorporates multiple perspectives.	1. AnchorAid aims to provide accessible, personalized, and early support for individuals experiencing post-traumatic stress symptoms, supplementing current mental healthcare providers. 2. We will conduct participatory design with care providers, recovered patients, research organizations, and policy-makers to incorporate their input into the design of AnchorAid.

Continued on next page

Table 7 – continued from previous page

Design cycle phase	Guidelines	AnchorAid Guidelines
Acquisition of Data and its Challenge	1. Identify and obtain a diverse set of data variables relevant to the mental health issue, including demographic information, clinical history, and psychosocial context. Use multiple data sources where feasible. 2. Address data limitations proactively. Build strategies to handle missing or biased data and ensure data quality and representativeness.	1. We will utilize EHRs and natural language to capture demographic and psychosocial predictors. We also consider wearable sensors as a promising avenue for further advancements. 2. We will ensure the data used is representative of diverse populations, including those in low- and middle-income countries, where AnchorAid has significant potential.
Ethical and Legal Considerations in Data Collection and Use	1. Design robust protocols for data privacy and security, incorporating transparency measures so users are aware of how their data is used and protected. 2. Actively seek and mitigate potential biases in both data and models. Use methodologies to regularly audit system fairness.	1. AnchorAid will comply with data regulations such as GDPR (EU), PIPEDA (CA), and ADPPA (US), and ensure transparent communication of these policies to the users. 2. We will use metrics to describe the model's fairness and set up regular audits to ensure system integrity.

Continued on next page

Table 7 – continued from previous page

Design cycle phase	Guidelines	AnchorAid Guidelines
Design, Selection, and Interpretability of Models	1. Select an AI model prioritizing transparency and interpretability. Users should be able to comprehend the model's logic, thus instilling trust in the system. 2. Ensure the system interface is intuitive, user-friendly, and accessible, providing a positive user experience.	1. We will consider using proven and interpretable AI models such as SVMs and DTs, which are commonly used in PTSD prognosis prediction. While CNNs can excel in complex pattern recognition, their less optimal interpretability is a factor we'll account for in our model selection process. 2. Using a user-centered design approach, we will use clear, engaging visual elements, ensure easy navigation, and prioritize accessibility, taking into account factors such as computer literacy, device usage, cultural differences, and language diversity.
Model Evaluation, Risks, and Impact	1. Use a diverse set of evaluation metrics, encompassing statistical performance and qualitative impact, such as user satisfaction and real-world effectiveness. 2. Establish risk assessment mechanisms, preparing for various scenarios including false positives and negatives. Create strategies to mitigate identified risks.	1. We will conduct both internal (original dataset) and external validation (independent dataset). In addition to traditional metrics (accuracy, AUC-ROC, F1-Score), we will evaluate real-world effectiveness through clinical trials, measuring the impact on patient outcomes. 2. We will establish a risk management framework, enumerating potential risks and outlining corresponding mitigation strategies. We consider missing a person in need a major risk and will thus ensure the model has a higher false positive rate.

Continued on next page

Table 7 – continued from previous page

Design cycle phase	Guidelines	AnchorAid Guidelines
System Deployment, Use, and Maintenance	1. Create a comprehensive deployment plan considering the unique contexts and workflows where the system will be used. 2. Develop a routine maintenance protocol that checks for data drift, system performance, user satisfaction, and regulatory changes.	1. We will run a prevention campaign to increase mental health awareness and establish a contact line (phone calls, emails) for individuals to learn about AnchorAid. We will also ensure users can discuss virtual agent's recommendations with mental health professionals via this contact line. We will develop protocols to seamlessly integrate AnchorAid into clinicians' workflows and ensure smooth information transfer throughout the clinical process. 2. The system will undergo regular maintenance based on the data used (data drifts, new findings from scientific research), user satisfaction (feedback, appreciation), audit results, and changes in regulations (data privacy, AI governance).

Continued on next page

Table 7 – continued from previous page

Design cycle phase	Guidelines	AnchorAid Guidelines
Feedback and Improvement	1. Establish continuous channels for user feedback and monitor user interaction with the system to identify potential improvements. 2. Adopt a culture of iterative development, continuously refining the system based on quantitative performance metrics and qualitative user feedback.	1. We will establish channels for users to provide detailed feedback about AnchorAid's interface, its recommendations, and users' health status. Depending on privacy considerations, these channels should enable communication across various stages of the clinical process, involving care providers, AnchorAid, the patient, and the family. 2. We will refine the system based on performance metrics, clinical trials, user feedback, and communication from other instances involved in the process.

Table 7: Design Guidelines for Integrating AnchorAid, a Virtual Support Tool for Post-Trauma Recovery

7 Discussion

7.1 Implications of the Study

This research has centered on generating robust design guidelines for the incorporation of AI into mental healthcare, thereby offering a wealth of insights regarding the challenges intrinsic to this fusion. Such hurdles include stigma and healthcare disparities; the complexity of diagnosis and prognosis; limited access to expertise; and the inherent subjectivity and reliance on self-reports in mental healthcare. Simultaneously, it is crucial to recognize that these challenges also herald opportunities, as AI holds the potential to surmount issues such as stigma, limited access to expertise, and self-report reliance. Additionally, AI has the capacity to encapsulate the complexity and multidimensional nature of mental health disorders' causality. Therefore, AI's potential transformative roles range from facilitating early detection and intervention, enhancing medical decision-making, personalizing and adapting treatment strategies, and scaling mental healthcare accessibility. This underlines the necessity for well-constructed design guidelines that aptly address the mentioned challenges and leverage arising opportunities.

Existing principles and guidelines for AI in general healthcare—derived from scientific articles, community letters, AI-related legal proposals, and recent government reports—are also examined in this study. These guidelines cover vital aspects, such as fairness and accountability, transparency, interpretability, and explainability, clinical validation and efficacy, privacy and security, as well as human-centered design and interoperability. My analysis of PTSD prognosis prediction case studies reveals an inconsistent application of these guidelines, though it should be noted that the primary objective of these studies is to further scientific research rather than provide practical, ready-to-deploy solutions. Drawing on the identified challenges and opportunities, the existing principles for AI in general healthcare, and the adherence of recent PTSD prognosis prediction studies to these principles, the study underscores the importance of the following key topics for the development of AI design guidelines in mental healthcare:

1. **Addressing Stigma and Access to Care**
2. **Enhancing Trust and Evaluate Outcomes**
3. **Understanding Heterogeneity and Complexity of Mental Health Disorders**
4. **Prioritizing Sensitivity to Privacy Concerns**
5. **Advocating for Patient-centered Care and Multidisciplinary Approaches**

These topics encompass crucial aspects that require attention during the design of AI systems for mental healthcare, based on a wide array of sources. The implications for the development of future design guidelines for AI systems in mental healthcare are substantial.

The study also illuminates the alignment of multidisciplinary collaboration, human-centered design, and AI system development. It underscores the benefits of such methodologies including a comprehensive understanding of user needs, AI systems designed to align with clinicians' workflow, facilitated shared decision-making, continuous system improvement through feedback loops, robust policymaking and ethical considerations, and equitable recognition of healthcare professionals' contributions. For a successful integration of multidisciplinary collaboration and human-centered design into AI system development, several key strategies are essential. These include the establishment of multidisciplinary teams and early, consistent involvement of end-users throughout the design process. Furthermore, fostering a culture characterized by transparency and trust, conscientiously addressing political and ethical considerations, and committing resources to continual evaluation and usability testing are crucial components of this integration.

Further, the study provides a comprehensive review of PTSD including its prevalence rates, diagnosis, prognosis, and treatment strategies. It also evaluates current machine and deep learning methods relevant to PTSD prognosis prediction and offers a summary of the latest applications of AI in PTSD prediction. This review can be instrumental for future studies merging AI and PTSD, especially given the scarcity of articles that explore recent advances in the field of PTSD prognosis prediction using AI systems.

Notably, the study culminates in the development of design guidelines for integrating AI into mental healthcare. These guidelines, derived from a thorough literature review and clinician and designer interviews, address the acknowledged need for assistance in the mental health sector due to the system's current strain. Clinicians and designers contributed valuable insights into the data, the models, the ethical needs, the risks, and the potential applications of such systems. These conversations gave rise to the concept of an AI tool that can aid both patients and care providers in managing PTSD. The guidelines, addressed to AI system designers, function as a checklist to ensure the safe and effective design and deployment of AI systems, taking into consideration the identified challenges and opportunities. They follow an innovative approach to AI system design, encompassing:

1. **Problem Definition and Stakeholder Engagement**
2. **Data Acquisition and Its Challenges**
3. **Ethical and Legal Considerations in Data Collection and Use**
4. **Design, Selection, and Interpretability of Models**

5. **Evaluation of Model, Risks, and Impact**
6. **System Deployment, Use, and Maintenance**
7. **Feedback and Improvement**

The resulting guidelines have considerable implications for future studies concerning AI and mental healthcare. They not only acknowledge and address the major risks associated with AI technologies but also facilitate the design of systems that best meet the needs of care providers, patients, and the overall mental healthcare system.

Finally, the study conceptualizes a virtual assistant — AnchorAid — that provides support for post-trauma recovery. The aim is not to supplant care providers but to offer early personal support to users and assist clinicians throughout the patient management process. AnchorAid also illustrates how the design guidelines I developed can be practically applied to manage post-trauma patient care.

7.2 Strengths and Limitations of the Research

This study stands out for its innovation and relevance. As AI increasingly garners attention, highlighted by the Future of Life Institute's recent call to halt the training of powerful AI systems, this study directly addresses a significant, timely concern. It also responds to recent government reports advocating for a comprehensive framework in AI design, further emphasizing its relevance to ongoing dialogues. Moreover, this study is underpinned by a recognized need for novel solutions in mental healthcare, underscoring its timely importance. In conducting this study, I worked closely with a PTSD research lab. This collaboration proved invaluable, providing unique insights and facilitating access to a global network of contacts. This direct experience brought depth and authenticity to the research. The study's broad, encompassing approach is another asset. By addressing a wide spectrum of issues across various stages of AI system design, this research offers a thorough exploration of the topic and a comprehensive understanding of the design process.

However, this research is not without its challenges. The novelty of the study's focus means there is limited literature specifically exploring AI in mental health and PTSD prediction. Some reports and letters I relied on for this study were published only recently, in 2023. This required continuous updating and adaptation of the findings, underscoring the need for the type of research I'm doing but making the process more challenging. The complexity of mental health disorders, with their inherent subjectivity, presents another significant challenge. Given the varied approaches the clinicians adopt toward stress management and treatment, and their differing opinions about AI in healthcare, these issues amplify the study's complexity. A logistical limitation was the conduct of a restricted number of interviews; only 10 were carried out due to time and resource constraints. Additionally, the initial intention was to analyze interviews based on automatic transcription and coding, however, a lack of resources and the hybrid format of interviews, some being online and others in-person, posed further challenges to the process. Finally, while the broad scope of this research is an asset, it also introduces difficulties. The wide array of information, spread across many articles, reports, letters, and interviews, makes it challenging to conduct a comprehensive analysis. No overarching review exists that covers all aspects of the topic.

In summary, this research has clear strengths, including its timely relevance, comprehensive approach, and privileged access to specialized resources. Yet, it's also confronted with challenges such as a limited body of related literature, the inherent complexity of the topic, the range of clinical opinions, and the difficulties associated with the broad scope of the study, along with logistical constraints in data collection and analysis.

7.3 Future Research Directions

The primary advancement necessary for the future of AI in mental healthcare involves the implementation of these systems within real-world scenarios and the rigorous evaluation of their impact through comprehensive clinical trials. Technically, we have come a long way, but we still need to address several crucial areas, including ethical considerations, bias mitigation, and assessment of clinical acceptability. The pressing need for these advancements stems from the direct implications AI systems have on people's lives, emphasizing the necessity of their effectiveness and ethical integrity.

Another significant aspect that warrants future research is the interpretability of AI systems. With deep learning technologies possessing the potential to analyze complex data such as speech and text, the '"black-box" nature of these models presents a notable challenge, particularly in clinical settings where transparency is paramount. Therefore, developing methods to increase interpretability or devising inherently more interpretable models will be instrumental in facilitating the adoption of these technologies.

Furthermore, future research should proactively adapt to the dynamically evolving landscape of AI governance. As new regulations and recommendations emerge, the design and deployment of AI systems should adjust accordingly to align with these ethical and legal frameworks. This foresighted adaptation during the design phase will ensure legal compliance and ethical alignment of the AI systems. Given that mental health patients may have trust concerns, a robust handling of data privacy is paramount in this sector, further underscoring the need for meticulous attention to data privacy in future research.

Additionally, a pivotal future research direction is enhanced communication and transparency about the purpose of AI tools. This study identified a prevalent concern among clinicians that they could be replaced by AI. It is important to stress that the goal of AI tools is not to replace, but to assist healthcare professionals. Further research should focus on developing strategies to alleviate these concerns, possibly through clearer communication and the involvement of clinicians in the design process.

Lastly, the future of this research field should also encompass a broader engagement with clinicians. Though this study has benefited from interviews with clinicians, further exploration was limited due to time and resource constraints. Most current articles and studies do not include care providers and patients enough in the design process of AI systems. Garnering more insights from clinicians and patients, who are the primary users of these AI systems, will significantly enhance their utility and adoption.

To summarize, future research should prioritize real-world applications and clinical trials, improve AI interpretability, adapt to evolving AI regulations, ensure data privacy, be transparent about the purpose of AI tools, and seek to include more clinicians in the research and development process. These concerted efforts will pave the way for realizing the full potential of AI in mental healthcare.

8 Conclusion

In conclusion, through my study titled "Design Guidelines for Integrating AI into Mental Healthcare: a Case Study on PTSD Prognosis Prediction", I aimed to create a practical pathway for embedding AI solutions in the context of mental health care. Recognizing the unique challenges and intrinsic complexities that characterize the mental healthcare domain, I brought to light the potential of AI to address these issues effectively.

I focused primarily on developing robust design guidelines for AI applications in mental healthcare. By conducting a comprehensive review of existing literature, examining PTSD prognosis prediction case studies, and gathering inputs from professionals, I was able to establish a suite of guidelines. These guidelines serve as a roadmap to address the nuances of mental health care, such as stigma, trust, privacy, and the need for patient-centered care. One significant outcome of my research is the hypothetical concept of AnchorAid, a virtual assistant designed to provide early support for individuals navigating post-trauma recovery and assist clinicians in managing PTSD. I consider AnchorAid as a practical embodiment of the developed guidelines, emphasizing their immediate applicability and real-world relevance. In the course of my research, I underscored the essential role of multidisciplinary collaboration and user-centered design in creating effective AI systems. The guidelines I formulated stressed the necessity for AI systems to be responsive to user needs while ensuring ethical and transparent operation.

Recognizing the limitations of my study, including the nascent nature of AI in mental healthcare and certain logistical constraints, I believe that its merits lie in its timeliness, comprehensive approach, and unique collaboration with a PTSD research lab. These elements solidify its substantial contribution to the field.

Looking ahead, I believe that potential directions for future research include rigorous evaluation of AI systems through real-world implementations and clinical trials, enhancing the interpretability of AI systems, aligning with evolving AI regulations, ensuring stringent data privacy, and improving transparency about the purpose of AI tools. I recommend that future research initiatives further involve clinicians and patients in the design and research process, ensuring that the AI tools developed are user-centric and clinically relevant.

To summarize, I believe my study significantly advances our understanding and provides practical guidelines for the integration of AI into mental healthcare. My development of these guidelines and the conceptualization of AnchorAid set the stage for the successful, ethically conscious deployment of AI systems in the mental health sector. This establishes a groundwork for a future where mental healthcare transcends geographical and socio-economic barriers, tailors uniquely to individual complexities, and leverages streamlined processes to significantly enhance patient outcomes.

9 Acknowledgments

I would like to extend my deepest appreciation to those who provided invaluable guidance and support throughout my Master's thesis journey. This work was pursued during my second year of the Master's program in Complex Systems Engineering and Management — CoSEM — at the TU Delft, under the expert supervision of esteemed professionals across two countries.

My heartfelt gratitude is extended to my primary supervisors at TU Delft, Nadia Metoui and Caroline Figueroa. Their unwavering support and insightful feedback, evidenced through our regular online meetings, were instrumental in refining the topic and structuring a coherent narrative for this thesis. They adeptly guided me in navigating the complexities of the research landscape, which greatly shaped the trajectory of my research work.

Kateřina Staňková, my chair at TU Delft, deserves special mention for her invaluable feedback at critical junctures. Her sharp observations significantly influenced the direction and scope of my research.

My academic journey was further enriched during my time at Mines ParisTech, where I participated in an Erasmus exchange program on the Bioengineering and Innovation in Neuroscience track. This experience deepened my interest in exploring the role of AI in mental health disorders, a quest substantially fostered by the continuous support of Nadia Metoui and Caroline Figueroa.

A significant part of my gratitude extends to Alain Brunet, my supervisor at the Douglas Research Center, affiliated with McGill University in Canada, where I undertook the bulk of my thesis work. His deep understanding of PTSD and his ceaseless willingness to engage in constructive debates and discussions provided invaluable insights into both the psychological and AI components of my research. His guidance was instrumental in my work, providing vital local supervision which complemented the remote guidance I received from my peers and mentors at TU Delft.

At the Douglas lab, Daniel Saumier's contributions significantly enriched my work, and I am grateful for his guidance. The entire Douglas lab team deserves my heartfelt appreciation. Their collaborative spirit during our general meetings, coupled with their diverse insights and feedback, was crucial in helping me identify promising research avenues and navigate potential challenges.

Finally, I wish to acknowledge the foundational training I received at the Brussels School of Engineering. This early education, while not directly related to this thesis, laid the groundwork for my academic journey by equipping me with essential skills in engineering, writing, and project management.

To all those who have contributed to my work in innumerable ways, my heartfelt thanks. Your contributions were instrumental to the successful completion of this project.

References

Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018, January). Big healthcare data: preserving security and privacy. *Journal of Big Data*, *5*(1), 1. Retrieved 2023-05-01, from https://doi.org/10.1186/s40537-017-0110-7 doi: 10.1186/s40537-017-0110-7

The Act. (2021, February). Retrieved 2023-05-09, from https://artificialintelligenceact.eu/the-act/

Adams, W. C. (2015, October). Conducting Semi-Structured Interviews. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of Practical Program Evaluation* (pp. 492–505). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved 2023-03-09, from https://onlinelibrary.wiley.com/doi/10.1002/9781119171386.ch19 doi: 10.1002/9781119171386.ch19

Alonso, J., Codony, M., Kovess, V., Angermeyer, M. C., Katz, S. J., Haro, J. M., ... Brugha, T. S. (2007, April). Population level of unmet need for mental healthcare in Europe. *The British Journal of Psychiatry*, *190*(4), 299–306. Retrieved 2022-10-14, from https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/population-level-of-unmet-need-for-mental-healthcare-in-europe/9B956FA5FFB8B970703F3DBA0DC69B5F (Publisher: Cambridge University Press) doi: 10.1192/bjp.bp.106.022004

Alzahrani, A. S., Gay, V., Alturki, R., & AlGhamdi, M. J. (2021, December). Towards Understanding the Usability Attributes of AI-Enabled eHealth Mobile Applications. *Journal of Healthcare Engineering*, *2021*, 5313027. Retrieved 2023-05-08, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8714331/ doi: 10.1155/2021/5313027

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & the Precise4Q consortium. (2020, November). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. Retrieved 2023-03-15, from https://doi.org/10.1186/s12911-020-01332-6 doi: 10.1186/s12911-020-01332-6

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... Horvitz, E. (2019, May). Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Glasgow Scotland Uk: ACM. Retrieved 2023-04-26, from https://dl.acm.org/doi/10.1145/3290605.3300233 doi: 10.1145/3290605.3300233

Andalibi, N., & Flood, M. K. (2021, January). Considerations in Designing Digital Peer Support for Mental Health: Interview Study Among Users of a Digital Support System (Buddy Project). *JMIR Mental Health*, *8*(1), e21819. Retrieved 2023-03-02, from https://mental.jmir.org/2021/1/e21819 (Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada) doi: 10.2196/21819

Anderson, P. L., Zimand, E., Hodges, L. F., & Rothbaum, B. O. (2005). Cognitive behavioral therapy for public-speaking anxiety using virtual reality for exposure. *Depression and Anxiety*, *22*(3), 156–158. Retrieved 2023-06-06, from https://onlinelibrary.wiley.com/doi/abs/10.1002/da.20090 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.20090) doi: 10.1002/da.20090

Andrade, L. H., Alonso, J., Mneimneh, Z., Wells, J. E., Al-Hamzawi, A., Borges, G., ... Kessler, R. C. (2014, April). Barriers to mental health treatment: results from the WHO World

Mental Health surveys. *Psychological Medicine*, *44*(6), 1303–1317. Retrieved 2022-10-13, from https://www.cambridge.org/core/product/identifier/S0033291713001943/type/journal_article doi: 10.1017/S0033291713001943

Anhang Price, R., Elliott, M. N., Zaslavsky, A. M., Hays, R. D., Lehrman, W. G., Rybowski, L., ... Cleary, P. D. (2014, October). Examining the role of patient experience surveys in measuring health care quality. *Medical care research and review: MCRR*, *71*(5), 522–554. doi: 10.1177/1077558714541480

Antoniou, G., Papadakis, E., & Baryannis, G. (2022, May). Mental Health Diagnosis: A Case for Explainable Artificial Intelligence. *International Journal on Artificial Intelligence Tools*, *31*(03), 2241003. Retrieved 2023-05-01, from https://www.worldscientific.com/doi/abs/10.1142/S0218213022410032 (Publisher: World Scientific Publishing Co.) doi: 10.1142/S0218213022410032

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017, January). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, *45*, 49–59. doi: 10.1016/j.janxdis.2016.11.008

Association, A. P. (2013). Diagnostic and Statistical Manual of Mental Disorders. Retrieved 2022-12-27, from https://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596 doi: https://doi.org/10.1176/appi.books.9780890425596

Barnett, M. L., Ray, K. N., Souza, J., & Mehrotra, A. (2018, November). Trends in Telemedicine Use in a Large Commercially Insured Population, 2005-2017. *JAMA*, *320*(20), 2147–2149. Retrieved 2023-04-30, from https://doi.org/10.1001/jama.2018.12354 doi: 10.1001/jama.2018.12354

Barrett, M., Oborn, E., Orlikowski, W. J., & Yates, J. (2012, October). Reconfiguring Boundary Relations: Robotic Innovations in Pharmacy Work. *Organization Science*, *23*(5), 1448–1466. Retrieved 2023-03-14, from https://pubsonline.informs.org/doi/abs/10.1287/orsc.1100.0639 (Publisher: INFORMS) doi: 10.1287/orsc.1100.0639

Bates, D. W., Auerbach, A., Schulam, P., Wright, A., & Saria, S. (2020, June). Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence. *Annals of Internal Medicine*, *172*(11_Supplement), S137–S144. Retrieved 2023-05-02, from https://www.acpjournals.org/doi/full/10.7326/M19-0872 (Publisher: American College of Physicians) doi: 10.7326/M19-0872

Bauer, M., Glenn, T., Geddes, J., Gitlin, M., Grof, P., Kessing, L. V., ... Whybrow, P. C. (2020, January). Smartphones in mental health: a critical review of background issues, current status and future concerns. *International Journal of Bipolar Disorders*, *8*(1), 2. Retrieved 2023-05-01, from https://doi.org/10.1186/s40345-019-0164-x doi: 10.1186/s40345-019-0164-x

Baumer, E. P. (2017, December). Toward human-centered algorithm design. *Big Data & Society*, *4*(2), 2053951717718854. Retrieved 2023-05-08, from https://doi.org/10.1177/2053951717718854 (Publisher: SAGE Publications Ltd) doi: 10.1177/2053951717718854

Bertl, M., Metsallik, J., & Ross, P. (2022, August). A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. *Frontiers in Psychiatry*, *13*, 923613. Retrieved 2023-03-15, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9396247/ doi: 10.3389/fpsyt.2022.923613

Bijl, R. V., & Ravelli, A. (2000, April). Psychiatric morbidity, service use, and need for care in the general population: results of The Netherlands Mental Health Survey and Incidence

Study. *American Journal of Public Health*, *90*(4), 602–607. Retrieved 2022-10-25, from https://ajph.aphapublications.org/doi/full/10.2105/AJPH.90.4.602 doi: 10.2105/AJPH.90.4.602

Birnbaum, M. L., Rizvi, A. F., Correll, C. U., Kane, J. M., & Confino, J. (2017, August). Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders: Online resources and pathways to care. *Early Intervention in Psychiatry*, *11*(4), 290–295. Retrieved 2023-05-01, from https://onlinelibrary.wiley.com/doi/10.1111/eip.12237 doi: 10.1111/eip.12237

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, 25–60. Retrieved 2023-05-01, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/ doi: 10.1016/B978-0-12-818438-7.00002-2

Bonanno, G. A., Mancini, A. D., Horton, J. L., Powell, T. M., LeardMann, C. A., Boyko, E. J., … Team, F. t. M. C. S. (2012, April). Trajectories of trauma symptoms and resilience in deployed US military service members: Prospective cohort study. *The British Journal of Psychiatry*, *200*(4), 317–323. Retrieved 2023-05-10, from https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/trajectories-of-trauma-symptoms-and-resilience-in-deployed-us-military-service-members-prospective-cohort-study/9BB21C44B18EE8CE6939BF44D90DD9C5 (Publisher: Cambridge University Press) doi: 10.1192/bjp.bp.111.096552

Branch, L. S. (2019, June). *Consolidated federal laws of canada, Personal Information Protection and Electronic Documents Act.* Retrieved 2023-05-01, from https://laws-lois.justice.gc.ca/eng/acts/p-8.6/ (Last Modified: 2019-06-21)

Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*(1), 5–32. Retrieved 2023-05-16, from https://doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Breslau, N. (2001). Outcomes of Posttraumatic Stress Disorder. *J Clin Psychiatry.*

Bryant, R. A., Nickerson, A., Creamer, M., O'Donnell, M., Forbes, D., Galatzer-Levy, I., … Silove, D. (2015, May). Trajectory of post-traumatic stress following traumatic injury: 6-year follow-up. *The British Journal of Psychiatry*, *206*(5), 417–423. Retrieved 2023-05-10, from https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/trajectory-of-posttraumatic-stress-following-traumatic-injury-6year-followup/BF901F6CD762ED0666C6FABA6353332F (Publisher: Cambridge University Press) doi: 10.1192/bjp.bp.114.145516

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019, November). "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 104:1–104:24. Retrieved 2023-03-16, from https://dl.acm.org/doi/10.1145/3359206 doi: 10.1145/3359206

Campbell, A., & Choudhury, T. (2012, March). From Smart to Cognitive Phones. *IEEE Pervasive Computing*, *11*(3), 7–11. Retrieved 2023-05-01, from http://ieeexplore.ieee.org/document/6228595/ doi: 10.1109/MPRV.2012.41

Carstensen, A.-K., & Bernhard, J. (2019, March). Design science research – a powerful tool for improving methods in engineering education research. *European Journal of Engineering Education*, *44*(1-2), 85–102. Retrieved 2022-11-22, from https://doi.org/10.1080/03043797.2018.1498459 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03043797.2018.1498459) doi: 10.1080/03043797.2018.1498459

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020, September). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. Retrieved 2023-05-16, from `https://www.sciencedirect.com/science/article/pii/S0925231220307153` doi: 10.1016/j.neucom.2019.10.118

Charmaz, K. (2014). *Constructing Grounded Theory.* SAGE. (Google-Books-ID: v_GGAwAAQBAJ)

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (Vol. 95, pp. 11–16). Cham: Springer International Publishing. Retrieved 2023-05-01, from `http://link.springer.com/10.1007/978-3-030-12524-0_2` (Series Title: Intelligent Systems, Control and Automation: Science and Engineering) doi: 10.1007/978-3-030-12524-0_2

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., ... Corlett, P. R. (2016, March). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet. Psychiatry*, *3*(3), 243–250. doi: 10.1016/S2215-0366(15)00471-X

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019, February). Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA Journal of Ethics*, *21*(2), 167–179. Retrieved 2023-05-01, from `https://journalofethics.ama-assn.org/article/can-ai-help-reduce-disparities-general-medical-and-mental-health-care/2019-02` (Publisher: American Medical Association) doi: 10.1001/amajethics.2019.167

Chen, J. H., & Asch, S. M. (2017, June). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *The New England journal of medicine*, *376*(26), 2507–2509. Retrieved 2022-12-27, from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5953825/` doi: 10.1056/NEJMp1702071

Cherrington, M., Airehrour, D., Lu, J., Xu, Q., Cameron-Brown, D., & Dunn, I. (2020, November). Features of Human-Centred Algorithm Design. In *2020 30th International Telecommunication Networks and Applications Conference (ITNAC)* (pp. 1–6). Melbourne, VIC, Australia: IEEE. Retrieved 2023-04-26, from `https://ieeexplore.ieee.org/document/9315169/` doi: 10.1109/ITNAC50341.2020.9315169

Chun Tie, Y., Birks, M., & Francis, K. (2019, January). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, *7*, 2050312118822927. Retrieved 2023-03-02, from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6318722/` doi: 10.1177/2050312118822927

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., ... Thornicroft, G. (2015, January). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, *45*(1), 11–27. doi: 10.1017/S0033291714000129

Conway, M., & O'Connor, D. (2016, June). Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology*, *9*, 77–82. Retrieved 2023-05-01, from `https://linkinghub.elsevier.com/retrieve/pii/S2352250X16000063` doi: 10.1016/j.copsyc.2016.01.004

Corrigan, P. (2004, October). How stigma interferes with mental health care. *The American Psychologist*, *59*(7), 614–625. doi: 10.1037/0003-066X.59.7.614

Corrigan, P. W., Druss, B. G., & Perlick, D. A. (2014, October). The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *15*(2), 37–70. doi: 10.1177/1529100614531398

Corrigan, P. W., & Rao, D. (2012, August). On the Self-Stigma of Mental Illness: Stages, Disclosure, and Strategies for Change. *Canadian journal of psychiatry. Revue canadienne de psychiatrie*, *57*(8), 464–469. Retrieved 2023-04-29, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3610943/

Costa, E., Soares, A. L., & de Sousa, J. P. (2016). Situating Case Studies Within the Design Science Research Paradigm: An Instantiation for Collaborative Networks. In H. Afsarmanesh, L. M. Camarinha-Matos, & A. Lucas Soares (Eds.), *Collaboration in a Hyperconnected World* (pp. 531–544). Cham: Springer International Publishing. doi: 10.1007/978-3-319-45390-3_45

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018, October). Assessing and addressing algorithmic bias in practice. *Interactions*, *25*(6), 58–63. Retrieved 2023-05-08, from https://dl.acm.org/doi/10.1145/3278156 doi: 10.1145/3278156

Creswell, J. (2009, January). Research Design: Qualitative, Quantitative, and Mixed-Method Approaches..

Cridland, E. K., Jones, S. C., Caputi, P., & Magee, C. A. (2015, January). Qualitative research with families living with autism spectrum disorder: Recommendations for conducting semistructured interviews. *Journal of Intellectual & Developmental Disability*, *40*(1), 78–91. Retrieved 2023-03-02, from https://doi.org/10.3109/13668250.2014.964191 (Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/13668250.2014.964191) doi: 10.3109/13668250.2014.964191

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011, June). The case study approach. *BMC Medical Research Methodology*, *11*, 100. Retrieved 2023-01-15, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3141799/ doi: 10.1186/1471-2288-11-100

Davenport, T., & Kalakota, R. (2019, June). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. Retrieved 2023-05-01, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/ doi: 10.7861/futurehosp.6-2-94

Demerouti, E. (2022). Turn Digitalization and Automation to a Job Resource. *Applied Psychology*, *71*(4), 1205–1209. Retrieved 2023-03-14, from https://onlinelibrary.wiley.com/doi/abs/10.1111/apps.12270 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apps.12270) doi: 10.1111/apps.12270

De Silva, D., & Alahakoon, D. (2022, June). An artificial intelligence life cycle: From conception to production. *Patterns*, *3*(6), 100489. Retrieved 2023-04-26, from https://www.sciencedirect.com/science/article/pii/S2666389922000745 doi: 10.1016/j.patter.2022.100489

de Vries, G.-J., & Olff, M. (2009). The lifetime prevalence of traumatic events and posttraumatic stress disorder in the Netherlands. *Journal of Traumatic Stress*, *22*(4), 259–267. Retrieved 2022-10-13, from https://onlinelibrary.wiley.com/doi/abs/10.1002/jts.20429 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jts.20429) doi: 10.1002/jts.20429

Dicicco-Bloom, B., & Crabtree, B. F. (2006, April). The qualitative research interview. *Medical Education*, *40*(4), 314–321. doi: 10.1111/j.1365-2929.2006.02418.x

Dyer, K. F. W., Dorahy, M. J., Hamilton, G., Corry, M., Shannon, M., MacSherry, A., ... McElhill, B. (2009, October). Anger, aggression, and self-harm in PTSD and complex PTSD. *Journal of Clinical Psychology*, *65*(10), 1099–1114. Retrieved 2023-05-01, from https://onlinelibrary.wiley.com/doi/10.1002/jclp.20619 doi: 10.1002/jclp.20619

Dückers, M. L. A., Alisic, E., & Brewin, C. R. (2016, October). A vulnerability paradox in the cross-national prevalence of post-traumatic stress disorder. *The British Journal of Psychiatry*, *209*(4), 300–305. Retrieved 2022-10-20, from https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/vulnerability-paradox-in-the-crossnational-prevalence-of-posttraumatic-stress-disorder/44FD260C9AC21868E34F6FC3968421E2 (Publisher: Cambridge University Press) doi: 10.1192/bjp.bp.115.176628

Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., ... Taylor, J. (2018). Internet- and mobile-based psychological interventions: Applications, efficacy, and potential for improving mental health: A report of the EFPA E-Health Taskforce. *European Psychologist*, *23*, 167–187. (Place: Germany Publisher: Hogrefe Publishing) doi: 10.1027/1016-9040/a000318

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017, February). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. Retrieved 2023-03-17, from https://www.nature.com/articles/nature21056 (Number: 7639 Publisher: Nature Publishing Group) doi: 10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2019, January). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. Retrieved 2023-01-15, from http://www.nature.com/articles/s41591-018-0316-z doi: 10.1038/s41591-018-0316-z

Ethics guidelines for trustworthy AI | Shaping Europe's digital future. (2019, April). Retrieved 2023-03-15, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013). *Using deep learning to enhance cancer diagnosis and classification.* (Journal Abbreviation: Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare Publication Title: Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare)

Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017, October). The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry*, *16*(3), 287–298. Retrieved 2023-05-01, from https://onlinelibrary.wiley.com/doi/10.1002/wps.20472 doi: 10.1002/wps.20472

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017, June). 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. Retrieved 2023-04-29, from https://mental.jmir.org/2017/2/e19 (Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada) doi: 10.2196/mental.7785

Forcier, M. B., Gallois, H., Mullan, S., & Joly, Y. (2019, September). Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers? *Journal of Law and the Biosciences*, *6*(1), 317–335. Retrieved 2023-05-01, from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813940/` doi: 10.1093/jlb/lsz013

Galatzer-Levy, I., Bonanno, G., Bush, D., & LeDoux, J. (2013). Heterogeneity in threat extinction learning: substantive and methodological considerations for identifying individual difference in response to stress. *Frontiers in Behavioral Neuroscience*, *7*. Retrieved 2023-05-12, from `https://www.frontiersin.org/articles/10.3389/fnbeh.2013.00055`

Galatzer-Levy, I. R., Huang, S. H., & Bonanno, G. A. (2018, July). Trajectories of resilience and dysfunction following potential trauma: A review and statistical evaluation. *Clinical Psychology Review*, *63*, 41–55. Retrieved 2023-05-10, from `https://www.sciencedirect.com/science/article/pii/S0272735818300539` doi: 10.1016/j.cpr.2018.05.008

Galatzer-Levy, I. R., Karstoft, K.-I., Statnikov, A., & Shalev, A. Y. (2014, December). Quantitative forecasting of PTSD from early trauma responses: A Machine Learning application. *Journal of Psychiatric Research*, *59*, 68–76. Retrieved 2022-11-24, from `https://www.sciencedirect.com/science/article/pii/S002239561400260X` doi: 10.1016/j.jpsychires.2014.08.017

Galatzer-Levy, I. R., Ma, S., Statnikov, A., Yehuda, R., & Shalev, A. Y. (2017, March). Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. *Translational Psychiatry*, *7*(3), e1070–e1070. Retrieved 2023-05-04, from `https://www.nature.com/articles/tp201738` (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/tp.2017.38

Galatzer-Levy, I. R., Nickerson, A., Litz, B. T., & Marmar, C. R. (2013). Patterns of Lifetime Ptsd Comorbidity: A Latent Class Analysis. *Depression and Anxiety*, *30*(5), 489–496. Retrieved 2022-10-13, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/da.22048` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.22048) doi: 10.1002/da.22048

Galatzer-Levy, I. R., Ruggles, K., & Chen, Z. (2018). Data Science in the Research Domain Criteria Era: Relevance of Machine Learning to the Study of Stress Pathology, Recovery, and Resilience. *Chronic Stress (Thousand Oaks, Calif.)*, *2*, 2470547017747553. doi: 10.1177/2470547017747553

Galatzer-Levy, I. R., Steenkamp, M. M., Brown, A. D., Qian, M., Inslicht, S., Henn-Haase, C., ... Marmar, C. R. (2014, September). Cortisol response to an experimental stress paradigm prospectively predicts long-term distress and resilience trajectories in response to active police service. *Journal of Psychiatric Research*, *56*, 36–42. Retrieved 2023-05-12, from `https://www.sciencedirect.com/science/article/pii/S0022395614001332` doi: 10.1016/j.jpsychires.2014.04.020

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014, February). Development of the CAT-ANX: A Computerized Adaptive Test for Anxiety. *The American journal of psychiatry*, *171*(2), 187–194. Retrieved 2023-05-01, from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4052830/` doi: 10.1176/appi.ajp.2013.13020178

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008, March). Methods of data collection in qualitative research: interviews and focus groups. *British Dental Journal*, *204*(6), 291–295. Retrieved 2023-03-02, from `https://www.nature.com/articles/bdj.2008.192` (Number: 6 Publisher: Nature Publishing Group) doi: 10.1038/bdj.2008.192

Gregório, J., Reis, L., Peyroteo, M., Maia, M., Mira da Silva, M., & Lapão, L. V. (2021, December). The role of Design Science Research Methodology in developing pharmacy eHealth services. *Research in social & administrative pharmacy: RSAP*, *17*(12), 2089–2096. doi: 10.1016/j.sapharm.2021.05.016

Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K. D. (2012). *The SAGE Handbook of Interview Research: The Complexity of the Craft*. SAGE. (Google-Books-ID: VCFsZsvZdwkC)

Gundersen, T., & Bærøe, K. (2022, April). The Future Ethics of Artificial Intelligence in Medicine: Making Sense of Collaborative Models. *Science and Engineering Ethics*, *28*(2), 17. Retrieved 2023-03-14, from https://doi.org/10.1007/s11948-022-00369-2 doi: 10.1007/s11948-022-00369-2

Hansebout, R. R., Cornacchi, S. D., Haines, T., & Goldsmith, C. H. (2009, August). How to use an article about prognosis. *Canadian Journal of Surgery*, *52*(4), 328–336. Retrieved 2022-12-03, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2724829/

Harrison, R., Flood, D., & Duce, D. (2013, May). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science*, *1*(1), 1. Retrieved 2023-05-08, from https://doi.org/10.1186/2194-0827-1-1 doi: 10.1186/2194-0827-1-1

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019, January). The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, *25*(1), 30–36. Retrieved 2023-03-14, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6995276/ doi: 10.1038/s41591-018-0307-0

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., . . . Riley, R. D. (2013, February). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ*, *346*, e5595. Retrieved 2022-12-28, from https://www.bmj.com/content/346/bmj.e5595 (Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting) doi: 10.1136/bmj.e5595

Hevner, A. (2007, January). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*.

Hevner, A., R, A., March, S., T, S., Park, Park, J., . . . Sudha (2004, March). Design Science in Information Systems Research. *Management Information Systems Quarterly*, *28*, 75.

HM Government. (2021, September). *National AI Strategy* (Tech. Rep.). Retrieved 2023-04-17, from https://www.gov.uk/government/publications/national-ai-strategy

Hoge, C. W., Castro, C. A., Messer, S. C., McGurk, D., Cotting, D. I., & Koffman, R. L. (2004, July). Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *The New England Journal of Medicine*, *351*(1), 13–22. doi: 10.1056/NEJMoa040603

Hollis, C., Falconer, C. J., Martin, J. L., Whittington, C., Stockton, S., Glazebrook, C., & Davies, E. B. (2017a, April). Annual Research Review: Digital health interventions for children and young people with mental health problems - a systematic and meta-review. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(4), 474–503. doi: 10.1111/jcpp.12663

Hollis, C., Falconer, C. J., Martin, J. L., Whittington, C., Stockton, S., Glazebrook, C., & Davies, E. B. (2017b, April). Annual Research Review: Digital health interventions for children and young people with mental health problems - a systematic and meta-review.

Journal of Child Psychology and Psychiatry, *58*(4), 474–503. Retrieved 2023-05-01, from https://onlinelibrary.wiley.com/doi/10.1111/jcpp.12663 doi: 10.1111/jcpp.12663

Hudlicka, E. (2016, January). Chapter 4 - Virtual Affective Agents and Therapeutic Games. In D. D. Luxton (Ed.), *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 81–115). San Diego: Academic Press. Retrieved 2023-05-08, from https://www.sciencedirect.com/science/article/pii/B9780124202481000040 doi: 10.1016/B978-0-12-420248-1.00004-0

Immy, H. (2005). *Qualitative Research In Health Care*. McGraw-Hill Education (UK). (Google-Books-ID: qXAwqGCl9rMC)

Insel, T. R. (2014, April). The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *The American Journal of Psychiatry*, *171*(4), 395–397. doi: 10.1176/appi.ajp.2014.14020138

Jin, W., Carpendale, S., Fraser, S., Hamarneh, G., & Gromala, D. (2019). Bridging AI Developers and End Users: an End-User-Centred Explainable AI Taxonomy and Visual Vocabularies.. Retrieved 2023-05-02, from https://www.semanticscholar.org/paper/Bridging-AI-Developers-and-End-Users%3A-an-AI-and-Jin-Carpendale/a6f54b7672600ca50cb54273a1b2bfead30ff95b

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023, January). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, *6*(1), 1–7. Retrieved 2023-05-01, from https://www.nature.com/articles/s41746-023-00751-9 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41746-023-00751-9

Kaba, R., & Sooriakumaran, P. (2007, February). The evolution of the doctor-patient relationship. *International Journal of Surgery*, *5*(1), 57–65. Retrieved 2023-03-15, from https://www.sciencedirect.com/science/article/pii/S1743919106000094 doi: 10.1016/j.ijsu.2006.01.005

Kallio, H., Pietilä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, *72*(12), 2954–2965. Retrieved 2023-03-02, from https://onlinelibrary.wiley.com/doi/abs/10.1111/jan.13031 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jan.13031) doi: 10.1111/jan.13031

Karstoft, K.-I., Galatzer-Levy, I. R., Statnikov, A., Li, Z., Shalev, A. Y., & For members of the Jerusalem Trauma Outreach and Prevention Study (J-TOPS) group. (2015, March). Bridging a translational gap: using machine learning to improve the prediction of PTSD. *BMC Psychiatry*, *15*(1), 30. Retrieved 2023-05-12, from https://doi.org/10.1186/s12888-015-0399-8 doi: 10.1186/s12888-015-0399-8

Kazdin, A. E. (2017, January). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy*, *88*, 7–18. Retrieved 2023-05-08, from https://linkinghub.elsevier.com/retrieve/pii/S0005796716300997 doi: 10.1016/j.brat.2016.06.004

Kazdin, A. E., & Blase, S. L. (2011, January). Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(1), 21–37. doi: 10.1177/1745691610393527

Keane, P. A., & Topol, E. J. (2018, August). With an eye to AI and autonomous diagnosis. *NPJ Digital Medicine*, *1*, 40. Retrieved 2023-05-02, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6550235/ doi: 10.1038/s41746-018-0048-y

Kearns, M. C., Ressler, K. J., Zatzick, D., & Rothbaum, B. O. (2012, October). Early interventions for PTSD: a review. *Depression and anxiety*, *29*(10), 833–842. Retrieved 2023-05-04, from `https://europepmc.org/articles/PMC3665083` doi: 10.1002/da.21997

Khairat, S., Marc, D., Crosby, W., & Sanousi, A. A. (2018, April). Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Medical Informatics*, *6*(2), e8912. Retrieved 2023-03-16, from `https://medinform.jmir.org/2018/2/e24` (Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada) doi: 10.2196/medinform.8912

Kim, W., Kim, K. S., Lee, J. E., Noh, D.-Y., Kim, S.-W., Jung, Y. S., ... Park, R. W. (2012, June). Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. *Journal of Breast Cancer*, *15*(2), 230–238. Retrieved 2023-03-23, from `https://doi.org/10.4048/jbc.2012.15.2.230` doi: 10.4048/jbc.2012.15.2.230

Kinsley, H., & Kukieła, D. (2020). *Neural Networks from Scratch in Python Book*. Retrieved 2023-05-16, from `https://nnfs.io/`

Knipscheer, J., Sleijpen, M., Frank, L., de Graaf, R., Kleber, R., ten Have, M., & Dückers, M. (2020, March). Prevalence of Potentially Traumatic Events, Other Life Events and Subsequent Reactions Indicative for Posttraumatic Stress Disorder in the Netherlands: A General Population Study Based on the Trauma Screening Questionnaire. *International Journal of Environmental Research and Public Health*, *17*(5), 1725. Retrieved 2022-10-13, from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7084195/` doi: 10.3390/ijerph17051725

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015, January). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. Retrieved 2022-12-27, from `https://www.sciencedirect.com/science/article/pii/S2001037014000464` doi: 10.1016/j.csbj.2014.11.005

Kubota, K. J., Chen, J. A., & Little, M. A. (2016, September). Machine learning for large-scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. *Movement Disorders: Official Journal of the Movement Disorder Society*, *31*(9), 1314–1326. doi: 10.1002/mds.26693

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved 2023-05-16, from `https://www.nature.com/articles/nature14539` (Number: 7553 Publisher: Nature Publishing Group) doi: 10.1038/nature14539

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., ... Jeste, D. V. (2021, September). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(9), 856–864. Retrieved 2023-03-17, from `https://www.sciencedirect.com/science/article/pii/S245190222100046X` doi: 10.1016/j.bpsc.2021.02.001

Lowe, S. R., Ratanatharathorn, A., Lai, B. S., van der Mei, W., Barbano, A. C., Bryant, R. A., ... Kessler, R. C. (2021, May). Posttraumatic Stress Disorder Symptom Trajectories Within the First Year Following Emergency Department Admissions: Pooled Results from the International Consortium to Predict PTSD. *Psychological medicine*, *51*(7), 1129–1139. Retrieved 2023-05-10, from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8318129/` doi: 10.1017/S0033291719004008

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016, January). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, *8*, 1–10. Retrieved 2023-05-01, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4720168/ doi: 10.4137/BII.S31559

Luxton, D., Pruitt, L., & Osenbach, J. (2014, February). Best Practices for Remote Psychological Assessment via Telehealth Technologies. *Professional Psychology Research and Practice*, *45*. doi: 10.1037/a0034547

Luxton, D. D. (2014, October). Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, *45*(5), 332–339. Retrieved 2023-03-17, from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0034559 doi: 10.1037/a0034559

Luxton, D. D., Anderson, S. L., & Anderson, M. (2016). Ethical Issues and Artificial Intelligence Technologies in Behavioral and Mental Health Care. In *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 255–276). Elsevier. Retrieved 2023-05-08, from https://linkinghub.elsevier.com/retrieve/pii/B9780124202481000118 doi: 10.1016/B978-0-12-420248-1.00011-8

Malgaroli, M., & Schultebraucks, K. (2020, October). Artificial Intelligence and Posttraumatic Stress Disorder (PTSD). *European Psychologist*, *25*(4), 272–282. Retrieved 2023-05-01, from https://econtent.hogrefe.com/doi/abs/10.1027/1016-9040/a000423 (Publisher: Hogrefe Publishing) doi: 10.1027/1016-9040/a000423

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016, November). Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, *26*(13), 1753–1760. doi: 10.1177/1049732315617444

Mandl, K. D., & Kohane, I. S. (2012, June). Escaping the EHR Trap — The Future of Health IT. *New England Journal of Medicine*, *366*(24), 2240–2242. Retrieved 2023-05-02, from http://www.nejm.org/doi/abs/10.1056/NEJMp1203102 doi: 10.1056/NEJMp1203102

Mason, M. (2010, August). Sample Size and Saturation in PhD Studies Using Qualitative Interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *11*(3). Retrieved 2023-03-02, from https://www.qualitative-research.net/index.php/fqs/article/view/1428 (Number: 3) doi: 10.17169/fqs-11.3.1428

McDonald, A. D., Sasangohar, F., Jatav, A., & Rao, A. H. (2019, July). Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: A supervised machine learning approach. *IISE Transactions on Healthcare Systems Engineering*, *9*(3), 201–211. Retrieved 2023-05-15, from https://doi.org/10.1080/24725579.2019.1583703 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24725579.2019.1583703) doi: 10.1080/24725579.2019.1583703

McDonald, S. D., & Calhoun, P. S. (2010, December). The diagnostic accuracy of the PTSD Checklist: A critical review. *Clinical Psychology Review*, *30*(8), 976–987. Retrieved 2023-05-10, from https://www.sciencedirect.com/science/article/pii/S0272735810001066 doi: 10.1016/j.cpr.2010.06.012

McDougall, R. J. (2019, March). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. doi: 10.1136/medethics-2018-105118

Mentis, A.-F. A., Lee, D., & Roussos, P. (2023, April). Applications of artificial intelligencemachine learning for detection of stress: a critical overview. *Molecular Psychiatry*, 1–13. Retrieved 2023-05-01, from https://www.nature.com/articles/s41380-023-02047-6 (Publisher: Nature Publishing Group) doi: 10.1038/s41380-023-02047-6

Mills, J., Bonner, A., & Francis, K. (2006). Adopting a constructivist approach to grounded theory: Implications for research design. *International Journal of Nursing Practice*, *12*(1), 8–13. Retrieved 2023-03-01, from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1440-172X.2006.00543.x` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1440-172X.2006.00543.x) doi: 10.1111/j.1440-172X.2006.00543.x

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018, November). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. Retrieved 2023-03-27, from `https://doi.org/10.1093/bib/bbx044` doi: 10.1093/bib/bbx044

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016, December). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. Retrieved 2023-05-08, from `https://doi.org/10.1177/2053951716679679` (Publisher: SAGE Publications Ltd) doi: 10.1177/2053951716679679

Mittelstadt, B. D., & Floridi, L. (2016, April). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, *22*(2), 303–341. doi: 10.1007/s11948-015-9652-2

Mohr, D. C., Burns, M. N., Schueller, S. M., Clarke, G., & Klinkman, M. (2013, July). Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health. *General Hospital Psychiatry*, *35*(4), 332–338. Retrieved 2023-05-01, from `https://linkinghub.elsevier.com/retrieve/pii/S0163834313000698` doi: 10.1016/j.genhosppsych.2013.03.008

Mohr, D. C., Cheung, K., Schueller, S. M., Hendricks Brown, C., & Duan, N. (2013, October). Continuous evaluation of evolving behavioral intervention technologies. *American Journal of Preventive Medicine*, *45*(4), 517–523. doi: 10.1016/j.amepre.2013.06.006

Moldoveanu, M. (2019, March). Why AI Underperforms and What Companies Can Do About It. *Harvard Business Review*. Retrieved 2023-03-15, from `https://hbr.org/2019/03/why-ai-underperforms-and-what-companies-can-do-about-it` (Section: Technology and analytics)

Musen, M. A., Middleton, B., & Greenes, R. A. (2021). Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine* (pp. 795–840). Springer.

Mutasa, S., Sun, S., & Ha, R. (2020, September). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical Imaging*, *65*, 96–99. Retrieved 2023-05-02, from `https://www.sciencedirect.com/science/article/pii/S0899707120301376` doi: 10.1016/j.clinimag.2020.04.025

Naslund, J. A., Aschbrenner, K. A., Araya, R., Marsch, L. A., Unützer, J., Patel, V., & Bartels, S. J. (2017, June). Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. *The Lancet. Psychiatry*, *4*(6), 486–500. doi: 10.1016/S2215-0366(17)30096-2

Nehra, D., Herrera-Escobar, J., Al Rafai, S., Havens, J., Askari, R., Nitzschke, S., ... Haider, A. (2019, July). Resilience and long-term outcomes after trauma: An opportunity for early intervention? *Journal of Trauma and Acute Care Surgery*, *87*, 1. doi: 10.1097/TA.0000000000002442

Ochoa, J. G. D., Csiszár, O., & Schimper, T. (2021, June). Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks. *BMC Medical Informatics and Decision Making*, *21*(1), 186. Retrieved 2022-11-10, from https://doi.org/10.1186/s12911-021-01553-3 doi: 10.1186/s12911-021-01553-3

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014, May). Quantifying heterogeneity attributable to polythetic diagnostic criteria: theoretical framework and empirical application. *Journal of Abnormal Psychology*, *123*(2), 452–462. doi: 10.1037/a0036068

Onnela, J.-P., & Rauch, S. L. (2016, June). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, *41*(7), 1691–1696. Retrieved 2023-05-01, from https://www.nature.com/articles/npp20167 (Number: 7 Publisher: Nature Publishing Group) doi: 10.1038/npp.2016.7

Park, S. H., Choi, J., & Byeon, J.-S. (2021, March). Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. *Korean Journal of Radiology*, *22*(3), 442–453. Retrieved 2023-05-02, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7909857/ doi: 10.3348/kjr.2021.0048

Parker, S. K., & Grote, G. (2022). Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. *Applied Psychology*, *71*(4), 1171–1204. Retrieved 2023-03-14, from https://onlinelibrary.wiley.com/doi/abs/10.1111/apps.12241 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apps.12241) doi: 10.1111/apps.12241

Patel, V., Chisholm, D., Parikh, R., Charlson, F. J., Degenhardt, L., Dua, T., … DCP MNS Author Group (2016, April). Addressing the burden of mental, neurological, and substance use disorders: key messages from Disease Control Priorities, 3rd edition. *Lancet (London, England)*, *387*(10028), 1672–1685. doi: 10.1016/S0140-6736(15)00390-6

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., … UnÜtzer, J. (2018, October). The Lancet Commission on global mental health and sustainable development. *Lancet (London, England)*, *392*(10157), 1553–1598. doi: 10.1016/S0140-6736(18)31612-X

Pause Giant AI Experiments: An Open Letter. (2023, March). Retrieved 2023-04-05, from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007, December). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. Retrieved 2022-11-22, from https://doi.org/10.2753/MIS0742-1222240302 (Publisher: Routledge _eprint: https://doi.org/10.2753/MIS0742-1222240302) doi: 10.2753/MIS0742-1222240302

Philippe, T. J., Sikder, N., Jackson, A., Koblanski, M. E., Liow, E., Pilarinos, A., & Vasarhelyi, K. (2022, May). Digital Health Interventions for Delivery of Mental Health Care: Systematic and Comprehensive Meta-Review. *JMIR Mental Health*, *9*(5), e35159. Retrieved 2023-05-01, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9109782/ doi: 10.2196/35159

Pietrzak, R. H., Feder, A., Singh, R., Schechter, C. B., Bromet, E. J., Katz, C. L., … Southwick, S. M. (2014, January). Trajectories of PTSD risk and resilience in World Trade Center responders: an 8-year prospective cohort study. *Psychological Medicine*, *44*(1), 205–219. Retrieved 2023-05-10, from https://www.cambridge.org/core/journals/psychological-medicine/article/abs/trajectories-of-ptsd-risk-and-resilience-in-world-trade-center-responders-an-8year-prospective-cohort-study/64F888AFBC86EDE3F29657ABE45910A3 (Publisher: Cambridge University Press) doi: 10.1017/S0033291713000597

Raileanu, L. E., & Stoffel, K. (2004, May). Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93. Retrieved 2023-05-12, from `https://doi.org/10.1023/B:AMAI.0000018580.96245.c6` doi: 10.1023/B:AMAI.0000018580.96245.c6

Ranchordas, S. (2021, May). *Experimental Regulations for AI: Sandboxes for Morals and Mores* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-04-15, from `https://papers.ssrn.com/abstract=3839744` doi: 10.2139/ssrn.3839744

Regel, S., & Joseph, S. (2017). *Post-traumatic Stress.* Oxford University Press. (Google-Books-ID: Btr_DQAAQBAJ)

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). (2016, April). Retrieved 2023-04-17, from `http://data.europa.eu/eli/reg/2016/679/oj/eng` (Legislative Body: EP, CONSIL)

Rep. Pallone, F. (2022, December). *Text - H.R.8152 - 117th Congress (2021-2022): American Data Privacy and Protection Act* [legislation]. Retrieved 2023-05-09, from `http://www.congress.gov/` (Archive Location: 12/30/2022)

Rizzo, A., Shilling, R., Forbell, E., Scherer, S., Gratch, J., & Morency, L.-P. (2016, January). Chapter 3 - Autonomous Virtual Human Agents for Healthcare Information Support and Clinical Interviewing. In D. D. Luxton (Ed.), *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 53–79). San Diego: Academic Press. Retrieved 2023-05-08, from `https://www.sciencedirect.com/science/article/pii/B9780124202481000039` doi: 10.1016/B978-0-12-420248-1.00003-9

Rokach, L., & Maimon, O. (2005). Decision Trees. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). New York: Springer-Verlag. Retrieved 2023-05-16, from `http://link.springer.com/10.1007/0-387-25465-X_9` doi: 10.1007/0-387-25465-X_9

Rosen, G. M., & Lilienfeld, S. O. (2008, June). Posttraumatic stress disorder: An empirical evaluation of core assumptions. *Clinical Psychology Review*, *28*(5), 837–868. Retrieved 2023-05-01, from `https://linkinghub.elsevier.com/retrieve/pii/S0272735807002048` doi: 10.1016/j.cpr.2007.12.002

Ruggiero, K. J., Price, M., Adams, Z., Stauffacher, K., McCauley, J., Danielson, C. K., ... Resnick, H. S. (2015, September). Web Intervention for Adolescents Affected by Disaster: Population-Based Randomized Controlled Trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*(9), 709–717. Retrieved 2023-05-01, from `https://linkinghub.elsevier.com/retrieve/pii/S0890856715004335` doi: 10.1016/j.jaac.2015.07.001

Saxe, G. N., Ma, S., Ren, J., & Aliferis, C. (2017, July). Machine learning methods to predict child posttraumatic stress: a proof of concept study. *BMC Psychiatry*, *17*(1), 223. Retrieved 2023-05-12, from `https://doi.org/10.1186/s12888-017-1384-1` doi: 10.1186/s12888-017-1384-1

Saxena, S., Sharan, P., Garrido, M., & Saraceno, B. (2006, October). World Health Organization's Mental Health Atlas 2005:implications for policy development. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, *5*(3), 179–184.

Schueller, S. M., Hunter, J. F., Figueroa, C., & Aguilera, A. (2019, September). Use of Digital Mental Health for Marginalized and Underserved Populations. *Current Treatment Options in Psychiatry*, *6*(3), 243–255. Retrieved 2023-04-29, from https://escholarship.org/uc/item/7463f1tk doi: 10.1007/s40501-019-00181-z

Schultebraucks, K., & Chang, B. P. (2021, February). The opportunities and challenges of machine learning in the acute care setting for precision prevention of posttraumatic stress sequelae. *Experimental Neurology*, *336*, 113526. Retrieved 2023-05-01, from https://linkinghub.elsevier.com/retrieve/pii/S0014488620303575 doi: 10.1016/j.expneurol.2020.113526

Schultebraucks, K., Qian, M., Abu-Amara, D., Dean, K., Laska, E., Siegel, C., ... Marmar, C. R. (2021, September). Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: a machine-learning approach for analyzing multivariate predictors. *Molecular Psychiatry*, *26*(9), 5011–5022. Retrieved 2023-05-15, from https://www.nature.com/articles/s41380-020-0789-2 (Number: 9 Publisher: Nature Publishing Group) doi: 10.1038/s41380-020-0789-2

Schultebraucks, K., Shalev, A. Y., Michopoulos, V., Grudzen, C. R., Shin, S.-M., Stevens, J. S., ... Galatzer-Levy, I. R. (2020, July). A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. *Nature Medicine*, *26*(7), 1084–1088. Retrieved 2023-05-04, from https://www.nature.com/articles/s41591-020-0951-z (Number: 7 Publisher: Nature Publishing Group) doi: 10.1038/s41591-020-0951-z

Schultebraucks, K., Sijbrandij, M., Galatzer-Levy, I., Mouthaan, J., Olff, M., & van Zuiden, M. (2021, May). Forecasting individual risk for long-term Posttraumatic Stress Disorder in emergency medical settings using biomedical data: A machine learning multicenter cohort study. *Neurobiology of Stress*, *14*, 100297. Retrieved 2022-11-24, from https://www.sciencedirect.com/science/article/pii/S2352289521000059 doi: 10.1016/j.ynstr.2021.100297

Shafiei, S. B., Lone, Z., Elsayed, A. S., Hussein, A. A., & Guru, K. A. (2020, December). Identifying mental health status using deep neural network trained by visual metrics. *Translational Psychiatry*, *10*(1), 1–8. Retrieved 2023-03-17, from https://www.nature.com/articles/s41398-020-01117-5 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41398-020-01117-5

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019, July). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448. doi: 10.1017/S0033291719000151

Shilton, K. (2018, July). Values and Ethics in Human-Computer Interaction. *Foundations and Trends® in Human–Computer Interaction*, *12*(2), 107–171. Retrieved 2023-05-08, from https://www.nowpublishers.com/article/Details/HCI-073 (Publisher: Now Publishers, Inc.) doi: 10.1561/1100000073

Shin, D. (2020, October). User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565. Retrieved 2023-05-01, from https://doi.org/10.1080/08838151.2020.1843357 (Publisher: Routledge _eprint: https://doi.org/10.1080/08838151.2020.1843357) doi: 10.1080/08838151.2020.1843357

Shin, D., Zhong, B., & Biocca, F. A. (2020, June). Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management*, *52*, 102061. Retrieved 2023-05-01, from https://www.sciencedirect.com/science/article/pii/S0268401219314161 doi: 10.1016/j.ijinfomgt.2019.102061

Shiraishi, J., Li, Q., Appelbaum, D., & Doi, K. (2011, November). Computer-Aided Diagnosis and Artificial Intelligence in Clinical Imaging. *Seminars in Nuclear Medicine*, *41*(6), 449–462. Retrieved 2023-03-17, from https://www.sciencedirect.com/science/article/pii/S0001299811000742 doi: 10.1053/j.semnuclmed.2011.06.004

Shneiderman, B. (2011, September). Claiming success, charting the future: micro-HCI and macro-HCI. *Interactions*, *18*(5), 10–11. Retrieved 2023-05-08, from https://dl.acm.org/doi/10.1145/2008176.2008180 doi: 10.1145/2008176.2008180

Shortliffe, E. H., & Sepúlveda, M. J. (2018, December). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, *320*(21), 2199–2200. doi: 10.1001/jama.2018.17163

Simonsen, J. (2004, January). Participatory IT Design. Designing for Business and Workplace Realities. MIT press. *Bødker, K., F. Kensing and J. Simonsen (2004). Participatory IT Design. Designing for Business and Workplace Realities. MIT press.*. Retrieved 2023-05-08, from https://www.academia.edu/38351927/Participatory_IT_Design_Designing_for_Business_and_Workplace_Realities_MIT_press

Singh, R. P., Hom, G. L., Abramoff, M. D., Campbell, J. P., & Chiang, M. F. (2020, August). Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. *Translational Vision Science & Technology*, *9*(2), 45. Retrieved 2023-03-15, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7443115/ doi: 10.1167/tvst.9.2.45

Sittig, D. F., & Singh, H. (2010, October). A New Socio-technical Model for Studying Health Information Technology in Complex Adaptive Healthcare Systems. *Quality & safety in health care*, *19*(Suppl 3), i68–i74. Retrieved 2023-05-01, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3120130/ doi: 10.1136/qshc.2010.042085

Stokols, D., Misra, S., Moser, R. P., Hall, K. L., & Taylor, B. K. (2008, August). The Ecology of Team Science: Understanding Contextual Influences on Transdisciplinary Collaboration. *American Journal of Preventive Medicine*, *35*(2, Supplement), S96–S115. Retrieved 2023-05-08, from https://www.sciencedirect.com/science/article/pii/S0749379708004091 doi: 10.1016/j.amepre.2008.05.003

Stuart, R., & Peter, N. (2016). *Artificial Intelligence - A Modern Approach 3rd Ed.* Berkeley. Retrieved 2023-05-16, from https://repository.unimal.ac.id/1022/

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020, April). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, *10*(1), 1–26. Retrieved 2023-03-28, from https://www.nature.com/articles/s41398-020-0780-3 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41398-020-0780-3

Tabassi, E. (2023). *AI Risk Management Framework: AI RMF (1.0)* (Tech. Rep. No. error: NIST AI 100-1). Gaithersburg, MD: National Institute of Standards and Technology. Retrieved 2023-04-17, from https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf doi: 10.6028/NIST.AI.100-1

Tang, T., Lim, M. E., Mansfield, E., McLachlan, A., & Quan, S. D. (2018, February). Clinician user involvement in the real world: Designing an electronic tool to improve interprofessional communication and collaboration in a hospital setting. *International Journal of Medical Informatics*, *110*, 90–97. Retrieved 2023-03-14, from https://www.sciencedirect.com/science/article/pii/S1386505617304252 doi: 10.1016/j.ijmedinf.2017.11.011

Topol, E. J. (2019, January). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. Retrieved 2023-03-15, from https://www.nature.com/articles/s41591-018-0300-7 (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41591-018-0300-7

Torous, J., Firth, J., Huckvale, K., Larsen, M. E., Cosco, T. D., Carney, R., ... Christensen, H. (2018, August). The Emerging Imperative for a Consensus Approach Toward the Rating and Clinical Recommendation of Mental Health Apps. *Journal of Nervous & Mental Disease*, *206*(8), 662–666. Retrieved 2023-05-01, from https://journals.lww.com/00005053-201808000-00013 doi: 10.1097/NMD.0000000000000864

Tulk Jesso, S., Kelliher, A., Sanghavi, H., Martin, T., & Henrickson Parker, S. (2022). Inclusion of Clinicians in the Development and Evaluation of Clinical Artificial Intelligence Tools: A Systematic Literature Review. *Frontiers in Psychology*, *13*. Retrieved 2023-03-16, from https://www.frontiersin.org/articles/10.3389/fpsyg.2022.830345

Tursunbayeva, A., & Renkema, M. (2022, April). Artificial intelligence in health-care: implications for the job design of healthcare professionals. *Asia Pacific Journal of Human Resources*. doi: 10.1111/1744-7941.12325

U.S. Chamber of Commerce. (2023, March). *Artificial Intelligence Commission Report* (Tech. Rep.). Retrieved 2023-04-17, from https://www.uschamber.com/technology/artificial-intelligence-commission-report

von der Warth, R., Dams, J., Grochtdreis, T., & König, H.-H. (2020, May). Economic evaluations and cost analyses in posttraumatic stress disorder: a systematic review. *European Journal of Psychotraumatology*, *11*(1), 1753940. doi: 10.1080/20008198.2020.1753940

Wainberg, M. L., Scorza, P., Shultz, J. M., Helpman, L., Mootz, J. J., Johnson, K. A., ... Arbuckle, M. R. (2017, May). Challenges and Opportunities in Global Mental Health: a Research-to-Practice Perspective. *Current Psychiatry Reports*, *19*(5), 28. doi: 10.1007/s11920-017-0780-z

Wang, P. S., Berglund, P. A., Olfson, M., & Kessler, R. C. (2004, April). Delays in Initial Treatment Contact after First Onset of a Mental Disorder. *Health Services Research*, *39*(2), 393–416. Retrieved 2022-10-31, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1361014/ doi: 10.1111/j.1475-6773.2004.00234.x

Wang, X., Zhang, H. H., & Wu, Y. (2019, July). Multiclass Probability Estimation With Support Vector Machines. *Journal of Computational and Graphical Statistics*, *28*(3), 586–595. Retrieved 2023-05-16, from https://doi.org/10.1080/10618600.2019.1585260 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2019.1585260) doi: 10.1080/10618600.2019.1585260

Watkins, L. E., Sprang, K. R., & Rothbaum, B. O. (2018). Treating PTSD: A Review of Evidence-Based Psychotherapy Interventions. *Frontiers in Behavioral Neuroscience*, *12*. Retrieved 2022-10-13, from https://www.frontiersin.org/articles/10.3389/fnbeh.2018.00258

Weiser, S. (2019, April). *Building trust in human-centric AI* [Text]. Retrieved 2023-05-08, from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Wilcox, H. C., Storr, C. L., & Breslau, N. (2009, March). Posttraumatic Stress Disorder and Suicide Attempts in a Community Sample of Urban American Young Adults. *Archives of General Psychiatry*, *66*(3), 305–311. Retrieved 2022-10-14, from https://doi.org/10.1001/archgenpsychiatry.2008.557 doi: 10.1001/archgenpsychiatry.2008.557

Wshah, S., Skalka, C., & Price, M. (2019, July). Predicting Posttraumatic Stress Disorder Risk: A Machine Learning Approach. *JMIR Mental Health*, *6*(7), e13946. Retrieved 2023-05-04, from https://mental.jmir.org/2019/7/e13946 (Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada) doi: 10.2196/13946

Yang, Q., Steinfeld, A., & Zimmerman, J. (2019, May). Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). Retrieved 2023-03-16, from http://arxiv.org/abs/1904.09612 (arXiv:1904.09612 [cs]) doi: 10.1145/3290605.3300468

Zhang, H., & Singer, B. H. (2010). *Recursive Partitioning and Applications* (Vol. 0). New York, NY: Springer. Retrieved 2023-05-16, from https://link.springer.com/10.1007/978-1-4419-6824-1 doi: 10.1007/978-1-4419-6824-1

Appendix A Glossary

ANN: Artificial Neural Network
AUC: Area Under the Curve
CAD: Computer-Aided Design
CBT: Cognitive Behavioral Therapy
CCDS: Computer-based Clinician Decision Support
CDSS: Clinical Decision Support System
CNN: Convolution Neural Network
CPT: Cognitive Processing Therapy
CSR: Case Study Research
DL: Deep Learning
DNN: Deep Neural Network
DSR: Design Science Research
DST: Decision Support Tools
DT: Decision Tree
EHR: Electronic Medical Record
EMDR: Eye Movement Desensitization and Reprocessing
FACT: Flexible Assertive Community Treatment
FNN: Feedforward Neural Network
GDPR: General Data Protection Regulation
GT: Grounded Theory
HCA: Health Care Application
HCI: Human-Computer-Interaction
HRQOL: Health-Related Quality Of Life
ICT: Information and Communication Technologies
IPO: Input-Process-Output
IPOE: Input-Process-Output-Engage
LMICs: Low- and Middle-Income Countries
MDD: Major Depressive Disorder
ML: Machine Learning
NB: Naive Bayes
NLP: Natural Language Processing
PE: Prolonged Exposure
PIPEDA: Personal Information Protection and Electronic Documents Act
PTSD: Posttraumatic Stress Disorder
RF: Random Forest
ROC: Receiver Operating Characteristic
SQL: Structured Query Language
SSRI: Selective Serotonin Reuptake Inhibitors
SVM: Support Vector Machine
TF-CBT: Trauma-Focused-CBT
UASAD: User Acceptance and System Adaptation Design

Appendix B Semi-structured Interviews

B.1 Semi-structured Interview Guide Development

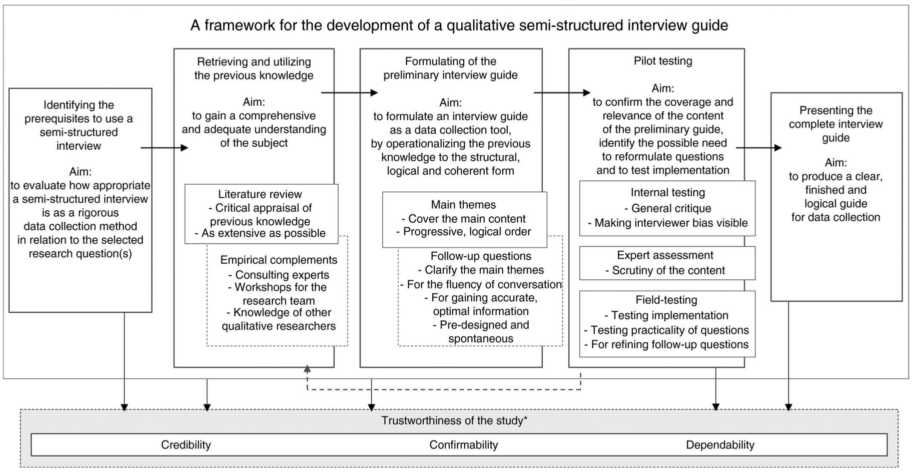

Figure 12: The phases of a semi-structured interview guide development (Kallio, Pietilä, Johnson, & Kangasniemi, 2016)

B.2 Grounded Theory Design Framework

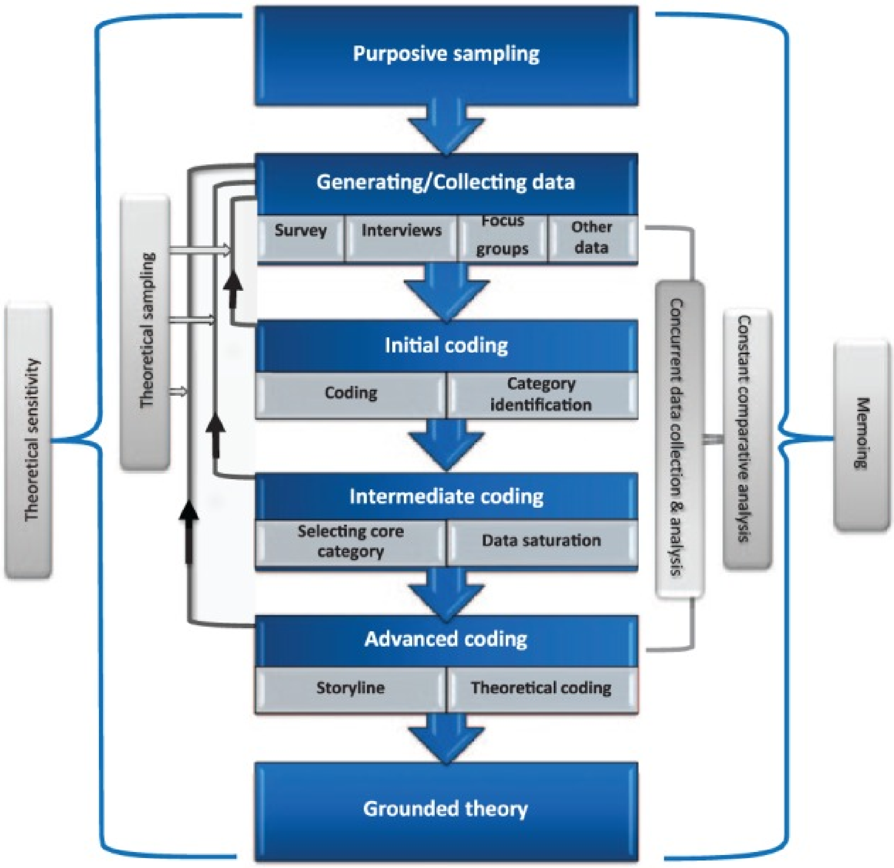

Figure 13: Grounded theory research - design framework (Chun Tie, Birks, & Francis, 2019)

B.3 Semi-structured Interview Guide

The following set of questions served as a guide for conducting interviews with clinicians, designers, and organizational directors. Depending on the participant's role and area of interest, I adapted these questions appropriately to elicit meaningful and relevant responses.

1. **Basic details**
   (a) Informant name, profession.

2. **Warm-up questions**
   (a) What does your day-to-day work consist of?
   (b) What do you know about PTSD and its treatment?
   (c) Have you ever heard about artificial intelligence, machine learning, or deep learning?
      i. What do you know about the application of artificial intelligence in mental health?
         A. And more globally in the medical field?
      ii. Have you ever been consulted for designing a technological tool, not especially in AI?
         A. How was your experience participating in the design of a learning tool for the medical field?

3. **First insights**
   Consider the following **case study**: "A deep learning-based tool is designed to predict the evolution of posttraumatic symptoms from identified key features such as patient's age, trauma, early signs, background, medical images, or treatment. It is based on an artificial neural network (ANN) made up of layers of interconnected 'neurons' that are trained on a dataset of past disease cases, along with associated factors such as the ones just mentioned. (ANNs belong to a sub-field of machine learning, called deep learning, inspired by the structure and function of the human brain. They learn to recognise patterns in data by adjusting the strengths of the connections between their neurons and can thus predict the likelihood of disease progression in new cases.)

   This method allows for processing vast amounts of data quickly and accurately and identifying patterns that might be missed by human clinicians. However, there are also limitations to using these tools, including the potential for bias if the dataset used to train them is not diverse or representative enough, the difficulty in interpreting their output, and the need for ongoing monitoring and validation.

   James, 73 years old, who participated in the Vietnam war comes to you, a psychiatrist specialised in PTSD because he often has violent dreams related to his trauma and he struggles to live his social life normally. The deep learning-based tool already successfully predicted the course of the disease in similar cases, and you want to use it for James's case to select the most appropriate treatment."

   Sofia, 26 years old, a student in medicine, and a victim of sexual assault comes to you, a psychologist specialised in PTSD, because she oRen has violent dreams related to her trauma and she struggles to meet and trust new people. The deep learning-based tool already successfully predicted the course of the disease in similar cases, and you want to use it for Sofia's case to select the most appropriate treatment."

   (a) What is your first impression of using such a tool?
      i. What aspect of your work do you think a machine learning-based tool could improve?
      ii. How do you imagine the workflow of using this tool?
      iii. What risks of using such a tool in mental healthcare do you identify?

4. **Specific questions**
   (a) Do you think that mental health clinicians and technological tool developers are connected enough?
      i. How this collaboration between the two could be improved?
      ii. As a clinician, what level of involvement would you like to have in the process of designing a machine-learning tool built for your clinical application?
         A. What could be the best format for collaborative design between designers, clinicians, and patients?
         B. How would you reach a team of designers if you have an idea for a tool that might improve your clinician's decision process?
         C. How much time per week, per month, would you spend meeting designers and helping them to build a tool promising for your work?
      iii. What is your opinion on integrating expert knowledge into the feature selection process?

A. How could you help select key data features meaningful for the prediction of the outcome desired?

iv. How would you like to be trained in using this tool?

(b) How would you use the results of the machine learning-based tool and incorporate them into your clinical expertise?

i. What evidence about the reliability of the tool would you need to use it?

ii. How explainable does the tool need to be?

iii. Would you communicate your process of choosing a treatment based on this tool to your patient?

iv. What could be the ethical risks associated with the use of these results?

A. Would you be concerned about data privacy?

5. **Final questions**

(a) What benefits of using machine learning-based tools in mental healthcare do you see?

(b) Overall, are you more concerned or enthusiastic about the use of this kind of tool?

i. Why?

(c) Did you learn anything new during this interview?

(d) Did it make you want to read more about artificial intelligence and its use in mental healthcare and more widely in the medical field?

B.4 Participant Information and Explicit Consent Points Forms

Study Title: Design Guidelines for Integrating AI into Mental Healthcare: A Case Study on Posttraumatic Stress Disorder Prediction

Researcher: Thomas Steurbaut, TU Delft, in collaboration with the Douglas Mental Health University Institute

Purpose of the Research: You are invited to participate in a research study aimed at understanding your perspectives and opinions about the use of AI technologies in mental healthcare practice. Your input, which should take approximately 45 minutes, will contribute to the development of design guidelines for effective implementation of deep learning tools in this field. The research will involve a case study on the use of deep learning and discussions about current processes, potential improvements, and future treatments.

What Will Happen: You will be asked to answer questions related to your professional views on the topic. If any question makes you uncomfortable, you are free to decline to answer.

Risks and Confidentiality: Although online activities inherently carry the risk of privacy breaches, we will take all necessary measures to protect your information. We are committed to maintaining confidentiality to the best of our ability. Your name and profession will not be included in the study. All data collection and analysis will adhere to the ethical guidelines of our university and current best practices, minimizing any potential risks.

Voluntary Participation: Participation in this study is entirely voluntary. You have the right to withdraw from the study at any time, without penalty.

Contacts:
Responsible Researcher: Nadia Metoui, email: N.Metoui@tudelft.nl
Researcher: Thomas Steurbaut, email: T.P.S.Steurbaut@student.tudelft.nl

If you have any further questions about the study or if you wish to withdraw your data, please do not hesitate to contact us.

Thank you for considering participating in this study. Your input is greatly valued.

Figure 14: Participant information

PLEASE TICK THE APPROPRIATE BOXES (also based on "Participant Information" statement)	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated 21/*03/2023 (date of redaction)*, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves:	☐	☐
Audio-recorded interview, transcribed as text		
4. I understand that I will be compensated for my participation by:	☐	☐
Not applicable		
5. I understand that the study will end:		
After +- 45 minutes		
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
6. I understand that taking part in the study involves the following risks […]. I understand that these will be mitigated by:	☐	☐
It will include professional views about the topic *Ability to stop the interview at any time, not answer the question, respect to right to be anonymous if desired*		
7. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) […] and associated personally identifiable research data (PIRD) […] with the potential risk of my identity being revealed:	☐	☐
Name, and profession will be the only personally identifiable information collected		
8. I understand that some of this PIRD is considered as sensitive data within GDPR legislation, specifically	☐	☐
Not applicable		
9. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach:	☐	☐
Anonymous data collection, or (pseudo-) anonymisation or aggregation if desired, secure data storage/limited access, transcription		
10. I understand that personal information collected about me that can identify me, such as my *name, age, profession*, will not be shared beyond the study team.	☐	☐
11. I understand that the (identifiable) personal data I provide will be destroyed:	☐	☐
When data are no longer needed for the research		
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
12. I understand that after the research study the de-identified information I provide will be used for:	☐	☐
Master's Thesis (Thomas Steurbaut, TU Delft – CoSEM) *Article publication (to be determined)*		
13. I agree that my responses, views, or other input can be quoted anonymously in research outputs	☐	☐
14. I agree that my real name can be used for quotes in research outputs	☐	☐
15. *If written information or other works are provided by the participants (e.g. in a reflection or other diary, or as images etc.) please check* https://www.tudelft.nl/library/copyright/c/what-is-copyright) *for information on copyright, and/or contact the Copyright Team for further information at* copyright-lib@tudelft.nl *and insert appropriate consent questions accordingly.* Not applicable	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
16. I give permission for the de-identified interviews transcribed that I provide to be archived in a personal repository so it can be used for future research and learning.	☐	☐
17. I understand that access to this repository is restricted only to the researcher, and not to the collaborating partner	☐	☐

Figure 15: Explicit consent points