Ethical and Practical Implications of Artificial Emotional Intelligence Approached from a Philosophical-Psychological Perspective- The Use Case of Psychiatry

Master's Thesis

Chiara Wüller

S1880357

University of Twente

Philosophy of Science, Technology and Society

1. **Supervisor: Anna Puzio**
2. **Supervisor: Julia Hermann**

Word count: 26.614

Submission Date: 22.09.2023

Abstract

Artificial Emotional Intelligence is one the visions for the future. It is promised to develop an AI that can recognise, interpret, process, and simulate emotions. This might enable more individual approaches and interactions, especially in the field of psychiatry. The aim of this research was to explore ethical and practical implications of AEI (in psychiatry). An interdisciplinary approach was used to assess the implications from a philosophical-psychological perspective. Five problem-centered expert interviews were conducted to ensure the practical applicability of the results and explore possible implications that might arise in the field of application. Practical implications of AEI in psychiatry among others include: improving the general mental health care system, the domain of application, privacy and data handling, which at the same time is an ethical implication. Ethical implications addressed lack of empathy and the uncanny valley problem, biases and responsibility, the need for governance, capitalisation, missing consensus on definitions of emotions, possible overreliance on the AI and the need for human oversight.

Table of Contents

Abstract ....

1. Introduction 1

2. Current State of Research 2

2.1 Affective Computing 3

2.2 AI and Mental Health 4

3. Method 7

4. Theoretical Concepts 8

4.1 Emotions 8

4.2 Emotional Intelligence 14

4.3 Artificial Emotional Intelligence 16

4.3.1 Artificial Intelligence 16

4.3.2 AI and Emotions 16

5. Problem-Centered Expert Interviews 18

5.1 Method of Expert Interviews 18

5.2 Interview Results 19

5.2.1 Interview Prof. Dr. Dr. Kai Vogeley 22

5.2.3 Interview Caroline Bollen 24

5.2.4 Interview Anna van Oosterzee 25

5.2.5 Interview Marloes Veldhuis 26

5.3. Analysis of the Expert-Interviews 28

6. Implications of AEI in Psychiatry 32

6.1 Ethical Considerations 32

6.2 Lack of Scientific Evidence and Transparency 39

6.3 Lack of Empathy and Compassionate Relationships 42

6.4 Uncanny Valley Problem 44

6.5 Examples of Current Developments and Applications 46

7. Discussion 50

7.1 Limitations of the Research 58

8. Conclusion 60

9. Literature 62

Appendix A 69

Appendix B 76

1. Introduction

Development of Artificial Intelligence (AI) systems is rapid and possibilities for application are tangent to diverse fields. AI systems become more complex and their performance increases and is promised to or already outperforms humans' performance in several areas (Asan et al., 2020, p. 4). AIs are promised to produce more accurate and individualized results, but despite all the promises, several problems and concerns arise. AI systems, especially those involving Deep Learning are so complex that humans are not always able anymore to explain or understand how the systems work (Janiesch et al., 2021, p. 692). Moreover, also the possibilities of AI are still limited, especially to computational tasks, hence AI based decision-making currently highly relies on pattern recognition, clustering, and categorization. This fact might imply that AI has its limits of application within tasks that might be more effective when other skills than computational skills are required. An example of this might be psychiatry. AI systems might be helpful when it comes to psychodiagnostics. They might be able to come up with more accurate diagnosis than human psychologists are able to, also because these tasks are often practiced by less experienced psychologists (Brown et al., 2019). However, when it comes to more abstract tasks, as psychotherapy, one could assume that AI would reach its limits. AI systems are tested also in therapeutical contexts and are meant to fulfil more abstract tasks, although they are currently still missing important required skills like abstraction and emotional intelligence (Booch et al., 2021, p. 15044).

Psychotherapy is a task that involves highly sensitive human interaction and a trust relationship between therapist and client (Slade & Holmes, 2019, p. 154). It is hard to imagine, how an AI system will be able to conduct the task sufficiently, when lacking the capability of being emotionally intelligent. Emotional Intelligence (EQ) is a skill that is already difficult for humans to achieve, so it is even harder to imagine that AI will achieve the state of being emotionally intelligent, when it is currently only reduced to computational intelligence. This raises severe concerns if AI systems should establish within psychiatric practice, when they are still lacking skills, as emotional intelligence that are needed to facilitate sensitive human interactions effectively. Hence, this work will address the possibilities of AI systems in acquiring the ability of being emotionally intelligent or at least acquiring necessary skills as preconditions to become emotionally intelligent (Ackley, 2016). Expertise from the disciplines of philosophy and psychology will be combined to illuminate the topic of Artificial Emotional Intelligence from different angles and explore interdisciplinary possibilities that might enable

AI development to move closer to Artificial Emotional Intelligence. Implications of possible Artificial Emotional Intelligence in psychiatry and psychotherapy will be explored.

Consequently, the research question this work aims to answer is "Can and should an Artificial Intelligence be emotionally intelligent especially within the field of psychiatry and what ethical and practical implications might an emotionally intelligent Artificial Intelligence have in psychiatry?". The thesis is that it is difficult for AI to learn the skill of being emotionally intelligent, however, if AI would be emotionally intelligent, it would be more acceptable to apply AI in psychiatry. Patients in psychotherapy are in a highly vulnerable condition, which makes it unethical to treat them with machines who cannot understand and validate their emotional states. Adding this skill to AI will make it more acceptable to apply AI in psychiatry but will come along with several ethical and practical implications, which currently still prevent application in the field. Adding the skill of Emotional Intelligence to AI systems might help in making the AI system more sensitive and its use in e.g., psychotherapeutic contexts more justifiable, but might also bring new ethical and practical implications to the table.

To answer the question better, I will first assess the current state of research within the relevant fields of affective computing and AI in mental health. Further I will introduce and define relevant theoretical concepts to set the stage for the following discussions. Next, I will introduce and analyse problem-centered expert-interviews, which I conducted to ensure the practical applicability of this research. After that I will discuss potential implications of Artificial Emotional Intelligence in psychiatry and end the section by giving some examples from current developments and applications. Then I will discuss my results and limitations of this research and end with summarizing the results, drawing a conclusion and provide an outlook for future research.

2. Current State of Research

Research in Artificial Intelligence (AI) is a hype currently, all kind of projects related to AI receive a lot of funding to explore the possibilities of AI and possible implications coming along with AI. The research fields this work is connected to closest are Affective Computing, which explores possibilities of combining Machine Learning (ML) approaches and emotions and the application of AI in mental health. Therefore, the current state of research within these fields will be presented in this section.

2.1 Affective Computing

The field of affective computing is a growing multidisciplinary field that relates to intentionally influencing emotions. It is concerned with questions that address how interactions between humans and technology can be facilitated by affect (Daily et al., 2017). If AI would be able to achieve emotional intelligence skills, it is promised that the technologies would better understand their users and improve interactions by being able to communicate more effectively with them (Czerwinski et al., 2021, p. 34).

Numerous complex models of emotions have been developed and there still is no common ground on what emotions really are (Barrett et al., 2019e). Psychology has likely offered the most accurate representation of emotions, however, only few models have been applied in the technical applications (Schuller & Schuller, 40). Two approaches, which either categorise emotions or understand them as continuous dimensions, serve as the basis for current emotion recognition and generation technologies. The basic-emotion-approach rooting to Paul Ekmann's work, which uses categorisation of emotions, and the problems of using this approach will be discussed in detail in the chapter about emotions. Preliminary, it is important to know, that he distinguishes six basic emotions. His approach was criticised for failing to capture the nuances, subjectivity, and context dependency of emotion. Nevertheless, the field of affective computing utilized basic emotion approaches to base AI systems on because these provide patterns to identify, which AI is capable of (Schuller & Schuller, 2018, p. 44). Affective computing in general aims to understand and replicate the dynamic nature of human emotions (Daily et al., 2017).

Wang et al. (2022, p. 19) argue that "affective computing is an umbrella term for human emotion, sentiment, and feelings, emotion recognition, and sentiment analysis computing". It was introduced by Professor Picard in 1997 with the aim to enable computers to intelligently facilitate and deal with human emotions. Various models of emotion recognition within the field of affective computing are based on psychological theories and metrics, while both deep learning (DL) and machine learning (ML) models are important for affective understanding (Wang et al., 2022, p. 42). Diverse types of databases are used for affective computing, including textual, audio, visual, physiological, and multimodal databases to ensure capturing all facets and roots of affective expression. While the field acknowledges that emotion recognition is a very difficult task, it promotes potential for technical agents, e.g., robots, being equipped with emotional intelligence to interact more effectively with humans and their environments.

So far, research on Artificial Emotional Intelligence was mainly focussing on developing tools for automated (human) emotion recognition and emotion generation in the domain of conversational agents and robots (Schuller & Schuller, 2018, p. 41). Along with the sparsity of consumer products comes the unawareness of products by the public. Technologies by now mainly target emotions which are “external” to AI, so they are externally observable, while AI’s “inner” emotions are not very complex yet. Text and haptic feedback, as well as synthesising emotional speech and facial expressions are underlying emotion recognition and generation. Contrary, attempts of active instrumentalization of emotions, as in emotion augmentation are rarely found in the literature. This would involve tasks where emotion is used for planning, reasoning, and more general goal achievement (Schuller & Schuller, 2018, p. 42). Nevertheless, there are still several unresolved technical challenges faced by engineers in the field of affective computing. These refer to the domains of multimodal natural language processing, affect detection, biometrics, and systems integration (Daily et al., 2017).

Philosophical literature has focussed mainly on possible ethical implications, which will be discussed in detail in a later chapter. Ethical considerations that are frequently addressed in literature about affective computing are privacy, transparency, bias and fairness, informed consent, emotional manipulation, security, and possible unknown other long-term implications (Daily et al., 2017). These implications are rooted in the field of affective computing, which envisions developing and deploying systems that can interact with human emotions. Grabowski et al. (2018, p. 59) made first attempts to use affective computing approaches and technologies to account for “emotion expression in psychiatric conditions”. However, they concluded, that “currently available methods have not been adequately validated in clinical settings”. They suggest, despite possible advantages in daily practice to further test and validate the affective computing tools before using them in real practice. Consequently, challenges this field is facing, are the missing consensus about what emotions are, the lack of existing AEI technologies and unsolved ethical concerns.

2.2 AI and Mental Health

In contrast to affective computing, literature addressing AI and mental health mainly focuses on assessing and discussing specific mental health applications. Since this work will focus specifically on the application in the psychiatric context, the current state of research will be presented mainly for this subdomain of mental health. However, it cannot be neglected that most mental health applications, which are already in use and are publicly available, are wellness and well-being applications, which can provide interesting insights for future psychiatric applications.

AI might have various possibilities of application in psychiatry. It could improve “planning of mental health services” and assist in “identifying and monitoring mental health problems” by using the “digitized health-care data” of a patient (WHO, 2023). Further, tasks can be automated by implementing AI applications, which can support clinicians in their work and help to move the understanding of the causes of more complex mental disorders to a deeper level, e.g.,, by leading and assessing the training of new clinicians (Abrams, 2023). AI could take over administrative tasks, as note taking during therapy sessions, highlighting themes and risks the clinician should review, to reduce clinician’s workload (Lovejoy et al., 2019, p. 2). Another possibility for AI systems to improve therapeutic practice is enabled by evidence-based protocols, by which the system can assess positive aspects and aspects which need improvement during the therapy session. It could detect the failure to ask key-questions and missed opportunities to validate the patient (Abrams, 2023). The company Lyssn was able to achieve a milestone, which would revolutionise Natural Language Processing (NLP), because their system is able to learn from unlabelled data, which enabled AI to transfer the learning to a specific target problem and to perform subtle but complicated tasks as detecting empathy (Allen, 2022). This enabled the quality monitoring and immediate feedback for trainees. Additionally, it is a first step in the direction of implementing emotional intelligent skills into AI systems.

One very common argument is that AI will replace many jobs, likely including psychiatry, since AI is able to make more rational decisions. The technology is promised to out-do humans regarding the mastery of diagnosis and treatment, which will be more individualised (Brown et al., 2021). Further, AI-led care will be more individualised as well. It is experienced as genuinely caring and understanding by the clients and AI agents have shown to successfully establish rapport with the clients to bond with them and built a trust relationship (Brown et al., 2021). This shows another move into the direction of emotionally intelligent AI. It is often claimed that humans tend to be more honest with technical agents than with real humans (Brown et al., 2021). Research has shown that it is easier for some patients to reveal their symptoms without fearing to be judged, which they perceive is given by AI (Ray et al., 2022, p. 4). Moreover, technologies are not influenced by natural human conditions as stress and fatigue and thereby could achieve better results in therapy (Ray et al., 2022, p.4). Additionally, less trails of different medication for one patient might be needed, because AI could calculate possible side effects and predictions (Lovejoy et al., 2019, p. 2).

Another prominent argument to implement AI in psychiatric practice, is its accessibility, lowered costs, and higher inclusiveness (Abrams, 2023; Ray et al., 2022). AI-based therapy, by e.g., chatbots, might provide an alternative for people, who are afraid of attending human-lead therapy or are new in therapy or suffer from social anxiety (Abrams, 2023). The vision is to make psychiatry more inclusive by building culturally capable AI. AI as a therapist could adapt features to match the patient's culture, e.g., dialect, to improve the patient's acceptance of the technology. Staff shortages, high costs and long wait lists are common factors preventing patients from receiving support (Hale, 2023). AI enabled- therapy could provide them cheaper and quicker access to support.

Moreover, mental illness still remains taboo in some areas of the world, AI could help patients to circumvent the resulting feelings of shame caused by looking for mental support in these areas. Further, since AI is not part of a wider social construct including cultural norms and expectations, it could overcome stigmatisation. AI is often seen as neutral, since it does not judge and has no opinion (Lovejoy et al., 2019, p. 2). On platforms like Reddit, users report about "their experience with the chatbot being as good or better than traditional therapy" (Hale, 2023). Hale (2023) at least sees potential for AI to be used to treat milder, well-known mental health conditions as depression and anxiety, where listening, validating, and offering practical steps the patient could take to address their problems, serve as standard treatment. Consequently, the constant availability of AI agents to assist and counsel clients is supposed to be a great benefit for clients during current times of shortage of professionals and health care systems preventing them to get immediate help. AI might be able to bridge this gap (Abbou, 2023).

Moreover, other stakeholders besides clinicians and patients can benefit from the use of AI. Through the data provided by AI systems and big data important insights to promote health by efficient strategies can be gathered by policymakers (WHO, 2023). The UK company ieso e.g., recorded online-therapy sessions over ten years, creating a rich dataset that could reveal what really works in therapy. AI can be used to analyse huge amounts of data (Allen, 2022). However, Hale (2023) suggests that AI in psychiatry might be most valuable in carrying out research and helping to assess patients progress, since AI's strength lies in pattern identification in data and making predictions and associations. Hence, AI as a data science, will help to rapidly screen large datasets and quick pattern recognition and faster disease detection (Ray et al., 2022;

p. 3). However, this would mean, that it might not be important for AI in psychiatry to achieve emotional skills, as its domain of use is not requiring them.

The Brown paper represents well the debate and tornness of the academic and psychiatric society. Considering the technological capability of current tools, on the one hand, is not a question if, but when the application of AI in psychiatry will be pursued. While on the other hand, a lot of scepticism, criticism and hesitance are prevailing. A detailed analysis of implications of AI in psychiatry will be provided in a later chapter. To sum up, an interdisciplinary approach will be needed to tackle the challenges and rule out the potential risks. Several fields need to combine their expertise to bring about the full potential of AI (Ray et al., 2022, p. 4).

Although one would expect to find research that combines affective computing or Artificial Emotional Intelligence approaches to the field of psychiatry, as emotions are a key element of human mental conditions and empathy plays an important role in psychotherapy, there is only little to none research to be found that combines these two fields. Only a few hints were found in the scientific literature regarding AI in psychiatry, that might imply attempts to move AI applications within this field to emotional capabilities. Hence, my work tries to combine the two AI domains and explores practical and ethical implications. Additionally, little was found in the popular philosophical literature and impactful philosophical journals. Often discussions only scratch ethical issues without going into depth, where they are rooted. The body of literature in psychological and AI research on the topic of AI and psychiatry in contrast is growing, but still relies hugely on speculations and visions than on real practical applications and implications. It was shown that currently most AI applications in psychiatry are not capable of emotional intelligent skills, so mainly they are simple AI applications. Hence, I will try to fill this gap by combining psychological and philosophical approaches on AI, emotions, and psychiatry. The benefits of my approach will be discussed in the next chapter.

3. Method

This research follows an interdisciplinary approach. It combines expertise from the fields of philosophy and psychology. Since the topic of Artificial Emotional Intelligence (AEI) is an interdisciplinary field itself, it is valuable to combine concepts from different disciplines to reach an accurate account of the topic. AEI, especially affective computing itself borrows its theoretical concepts and foundations mainly from the domain of psychology, because this discipline was claimed to provide the most accurate conceptualisation of e.g., emotions (Schuller & Schuller, 2018). However, since there is no clear consensus on the definitions, which makes it valuable to include philosophical perspectives to critically review the concepts. Additionally, philosophical questions, especially ethical issues are predominant concerns with the topic of AEI, hence philosophical concepts are helpful in assessing these. Additionally, the use case combines AI and psychiatry, which makes a theoretical foundation of the research in philosophical as well as psychological approaches even more important.

Psychiatry approaches mental health and psychological problems from the medical field, while psychology does mainly approach them from the behavioural field (de Bruin, 00:00:18). Hence, psychology provides data, empirical studies, and contexts, while philosophy adds theoretical and ethical components to the debate, helps to critically assess the data and contexts and combine them into a coherent picture. Calls for more interdisciplinary approaches for addressing societal problems in research become more frequent and the benefits of interdisciplinary knowledge exchange are valued more and more. Hence one of the strengths of this research is that it benefits from the different methods of knowledge production and transference of the two disciplines.

To set the stage and provide sufficient context for later discussions, in the following first some necessary concepts will be presented and defined that will play a crucial role in later discussions.

4. Theoretical Concepts

4.1 Emotions

Figuring out other people, their behaviour and understanding why they do what they do is always a difficult task. Humans try to sense other humans to establish relationships and be able to communicate and live along together as a society. Humans do not only express themselves via verbal expression, mimic and gestures are also parts of human communication, though they are the more obvious observable parts. A more complex and difficult part of human expression are their emotions. But what are human emotions and why are they so powerful?

In daily communication people often confuse feelings with emotions, however they need to be distinguished. While feelings are subjective bodily experiences, emotions are subconscious results of chemical or electrical processes that control human behaviour (Abbou, 2023, p. 53; Aday et al., 2017). Approached from the psychological perspective, feelings describe humans' innate ability to feel something bodily, like hungriness or satisfaction. They are something humans experience right in that moment. It is a bodily subjective experience, since one can only feel their own feelings within their body. If one tries to feel what others feel, one can only imagine how the other feels and the imagination is happening inside the brain, but there is no bodily experience of really feeling what the other feels. Hence, feelings are conscious because we can experience them. Contrary, emotions are subconscious. They influence learning and thought processes without us humans recognizing it. The explanations for these processes lie in the neural mechanisms of the brain (Aday et al., 2017). Emotions are goal-oriented, since they serve to processes and react to stimuli very quickly and effectively (Abbou, 2023, p. 54). They are very complex and consist of multiple components, which makes feelings only a small part of emotions.

However, approaching emotions from a philosophical perspective, one could come to a different conclusion. LeDoux and Brown (2017, p. E2020) for example, made efforts to explain emotions by a Higher-Order Theory (HOT) of Emotional Consciousness. By doing this, they explain that emotions are fundamentally tied to the self. Emotions depend on one's self-awareness and the ability to empathize with others. Without a sense of self, there would be no emotions like fear or happiness. LeDoux and Brown (2017) argue that emotional consciousness, as well as self-awareness and emotional self-awareness, are all linked to and controlled by a single neural circuit known as the General Neural Circuit (GNC). Emotional and non-emotional states of awareness share fundamental neural mechanisms but differ in the inputs they process. These inputs refer to the specific sensory and cognitive information that triggers the conscious experience. Although, this is a philosophical approach to emotions, it can be seen that the authors borrow arguments from the cognitive sciences to validate their statements. Hence, it is almost impossible to consider the psychological and philosophical perspectives distinctly.

The research field surrounding emotions is young and recently brings about diverse interesting insights that can be used in a diversity of fields. Paul Ekmann, an American psychologist, was the first to develop a strategy to measure emotions, make them graspable and usable for further research and come up with a classification of six, later seven, basic emotions: “Happiness, sadness, fear, anger, disgust, contempt and surprise” (Tracy & Randles, 2011, p. 399). Of course, there were many others coming up with their own classifications and definitions of emotions. Reisenzein (2007) believes that it is not even necessary to have a universal definition of emotion to be able to conduct meaningful research. He thinks it is not even possible to come up with a definition of emotion unless substantial empirical research is conducted. Until today there is still no consensus found for a unified definition of emotions. Every discipline has its own approaches and conceptualisations of emotion and often the disciplines borrow explanations from other disciplines, mostly from the cognitive sciences as psychology and neurosciences. Nevertheless, Ekmann's work figured out to remain essential and built the basis for the development of methods to measure emotions, which I will elaborate later.

Today there are several methods to measure emotions, however when the research line started, there were only limited measures of bodily reactions to emotions, serving as rough parameters, like heartrate, which often were too unspecific and similar (Abbou, 2023, p. 55). Newer technologies enabled more accurate measurements of emotions, as the functional magnetic resonance imaging (fMRI). FMRI is a method to visualise internal physical functions within the human body. This new technology revolutionised "methods to study neural processes and correlates of emotional behaviours" (Michalska & Davis, 2018, p. 417). Emotions can be visually represented by the blood flows within different brain areas (Abbou, 2023, p. 57). Different emotions stimulate activity in different brain areas, which can be observed on the fMRI scans. Chemical reactions stimulate electrical impulses, which activate the release of neurotransmitter, which either activate or inhibit neurons, which in the end results in a reaction of the human body. Neurotransmitters are among other hormones that active happiness, as endorphins, serotonins and oxytocins that evoke processes within the human bodies that are experienced as emotions (Butnariu & Sarac, 2019). Those chemical and electrical processes are measurable. Hence, emotions seem to be rooted within humans more rationally than initially expected.

However, emotions are still subjective and differ from human to human in intensity, expression, and interpretation. Some people experience and express their emotions more intensely and openly, while others try to hide their emotions as good as possible. Additionally, humans need to interpret emotions of other humans, which can be very difficult because of the subjective character of emotions (Barrett et al., 2019). They do not only vary between persons, but also among cultures and are highly context dependent. Facial expressions play a crucial role within this process. Emotions can be transferred to others via facial expressions by showing others the emotions, e.g., by smiling at them. Further, research has shown that people adapt to the facial expressions of others during interaction, which is a way of showing empathy (Abbou, 2023, p. 60). People mirror the emotions of others and thereby show understanding. Neuroscientist and psychologist Lisa Feldmann Barrett identified four criteria that need to be fulfilled to correctly interpret facial expressions: reliability, specificity, generalizability, and validity (Barrett et al., 2019, p. 9). Barrett and colleagues conclude that there is no universal interpretation of emotions, which poses a fundamental critic to Ekmann's classification approach.

Psychologist Andrew Ortony (2022, p. 41) state that there does not even exist a “stable criterion of basicness“ of emotions since several researchers have come up with “substantially different lists of basic emotions”. He argues that the issue lies within relying heavily on affective terms to label emotions, while the so-called basic emotions do not even defy certain criteria Ortony considers requirements of emotions, which he elaborate on the example of surprise (2022). Hence, he concludes that the basic emotions have no validation and calls for more substantial definitions for emotions.

Barrett and colleagues (2019, p. 6) acknowledge Ekmann’s work to have been very influential for research on emotion during the past 50 years. However, they claim that the common view, how they frame the dominant research line established based on Ekmann’s work among others that persons’ emotional states are revealed in facial movements, cannot be supported by the existing scientific evidence anymore (Barrett et al., 2019; Ortony, 2022). Still, the common view in omnipresent and serves as a basis for a lot of research, although the same scientists, relying on the basic-emotion approach, acknowledge that there is variation in expression and interpretation of emotions. Barrett and colleagues warn consumers of such scientific literature to be aware of these contradictory claims within the literature. The researcher group around Barrett also admits that they have relied on the common view in past publications and as a basis for the studies they conducted, “ignoring the importance of individual and contextual variation”, since the common view was dominant in scientific literature during that time (2019, p. 6). The common view happens to persist despite being refuted and found its way into other disciplines concerned with understanding emotions, as AI research, which I will explain later.

Problematic is that the common view powerfully influenced and shaped parts of society. The assumption underlying the common view guides “diagnosis and treatment of psychiatric illness” and it seems as if “at least in certain parts of the world, people believe that certain emotion categories are reliably signaled or revealed by certain facial-muscle movement configurations“(Barrett, et al., 2019, p. 1; 3). Worrying is also that these unsound believes have manifested in cultural products, especially modern technologies. The best examples are emojis or emoticons. Almost everyone uses them within electronic messaging to express their emotions. However, the way these emojis or emoticons are designed is just another schematized version of the proposed facial expressions that are meant to represent various emotion categories. They are considered as being universal, despite weak scientific evidence. Further, the common view mainly is only valid in Western industrialized, urban areas.

Barrett et al. (2019) conducted an extensive literature review, to get an overview of the existing scientific evidence for the common view, focusing on the basic-emotion approach with the six basic emotions. They found overall weak to no valid or reliable evidence supporting the common view. Hereby, it was interesting to observe that evidence was higher when the studies of emotion recognition were conducted with pre-scribed labels. The evidence for the reliability of emotion inferences was much weaker when methods included reverse-correlation or free-labelling tasks (p. 37). This means that participants were more likely to provide the same labels for emotions recognised, when they had pre-scribed categories, while there was more diversity in labelling, when participants had to come up with the categories themselves. Further, they found evidence for cultural variation in emotion perception, but were also suggesting conducting more extensive research, especially with participants from remote communities, children, elderly people, and people who are congenitally blind to better understand the real nature of emotion perception.

There is a substantial variation in emotion expression "across cultures, situations and even across people within a single situation" (Barrett et al., 2019, p. 1). Abbou (2023, p. 62) visualised some of these variations from her own experience in her book. She is from Morocco and migrated to Germany, so she grew up with African habits, e.g., she explains that it is very common in Morocco to use a lot of gestures during conversation and touch the conversation partner to express proximity, while in Germany she had to hold herself back from touching strangers during conversations to avoid misunderstandings or even intimidating those persons. Further, in Morocco no one would raise their voice, while in Germany it is common to casually shout from one floor to the other in the house to tell your family that dinner is ready. When Abbou first encountered these incidents, she got anxious and thought something happened or felt intimidated, because she was not used to someone raising their voice casually. Hence, she had to learn how the different cultures express themselves and assess and interpret the situations accordingly.

This stresses the relevance of seeing emotion expression, especially the supposed related facial movements as varying. They are highly dependent on the "immediate context, which includes a person's internal context" and the "outward context", "both of which vary in dynamic ways over time" (Barrett, 2019, p.4). Further, no controlled studies were conducted across remote cultures yet, hence there is too little evidence to draw conclusions about emotion expression and recognition for remote cultures (ibid., p. 23). That emotion perception abilities emerge over time and are shaped by a learning process within a social environment, is shown by data obtained from infants and young children (ibid., p. 45). People's emotion categories in general are flexible and responsive dependent on the stimuli they are presented to in their individual environments. Facial expressions are not as generalisable and universal as the common view presents, since they can also be linked to non-emotional psychological meanings that match specific situations or cultural contexts. While the face is a powerful tool during social interactions, there are no unique facial movements that can be exclusively linked to an emotion. For example, a scowl can be *an* expression of anger, but is not exclusively *the* expression of anger (Barrett, 2023, p. 46).

Summarizing the main concerns Barrett and colleagues (2023) identified, there are "three key shortcomings in the scientific research" on emotions that reinforced the misconception about how emotions are expressed and perceived, especially regarding the role of facial movements. The first shortcoming is the limited reliability. There is no reliable evidence that instances of the same emotion category are expressed or perceived by a common set of facial movements (ibid., p.3). The second shortcoming addresses the lack of specificity, which relates to the fact that instances of emotion categories and facial movement configurations cannot uniquely be mapped. The last shortcoming tackles the limited generalizability, which means that effects of context and culture are only insufficiently documented and accounted for in the literature supporting the common view. They conclude that facial configurations, especially the basic emotions that were in focus of the review "are best thought of as Western gestures, symbols or stereotypes that fail to capture the rich variety with which people spontaneously move their faces to express emotions in everyday life" (Barrett et al., 2023, p. 46). Thereby, they highlight, the lack of emotion research of minorities and vulnerable groups and the lack of translation of the research findings onto real world experiences.

Another philosophical approach to emotions is presented by Catrin Misselhorn, which she bases on the theory of the philosopher and psychologist Willian James. According to Misselhorn (2021, p. 14) emotions can be defined as episodic reactions with a specific beginning and ending. The subjective experiential quality of emotions builds the phenomenological consciousness. Emotions relate to an object, which is intentionality according to philosophical emotion theory, consequently, the reference objects are intentional objects, which are not necessarily of material nature. Intentional in this context means that emotions refer to something, they are directed or about something, which is the intentional object (Shargel, 2016, p. 46). The way of personal appraisal of those objects evokes emotions. A positive personal appraisal of an object for example would elicit positively perceived emotions as happiness. That emotions are rather passive by nature, is a prove that humans do not provoke them on purpose (Misselhorn, 2021, p.15). Emotions manifest themselves in language, attitude, gestures, and mimicry and cause specific behaviours. The appraisal of emotions might assess factors as the relevance for oneself, the novelty, the hedonic quality, and probability of occurrence. The appraisal patterns might involve cognitive complex processes, which are dependent on culture and context. The subjectivity, culture- and context-dependency of emotions is in line with Barrett et al.'s claims for more diversity.

Despite the varieties in expressing and interpreting emotions, people can also regulate, control, and hide their emotions to a certain extent, e.g., to prevent others from knowing too much about them (Eisenberg & Spinrad, 2004). This requires the skill of emotional intelligence, which will be explained in the next chapter.

4.2 Emotional Intelligence

Emotional Intelligence (EQ or EI) is a concept that established because the scientific community was curious about other forms of intelligence. They believed there needed to be more ways to succeed in life than just relying on cognitive intelligence and IQ. However, as it is the case with cognitive intelligence, it is impossible to come up with a universal definition of EQ, which is shared and agreed on among the whole community (Birks & Watt, 2007, p. 368). There are various strains of research and practice surrounding EQ and what EQ is, highly depends on whom you ask (Ackley, 2016, p. 271). Ackley (2016) reviewed the three major strains by Mayer, Salovey, & Caruso (2002), Goleman (1995; 1998) and Bar-On (1997). She identified as similarities that EQ is identified as a set of skills and capabilities by all the three research lines. Further, there is a learning effect involved, the skills can either be acquired or are at least improvable. Here opinions diverge among the authors. Salovey and Mayer (1989 – 1990) believe that EQ, like IQ, is inborn and improvement of EQ is limited by the inherent abilities, which is why "their model is referred to as an ability model" (Ackeley, 2016, p. 271). Goleman (1995, 1998) and Bar-On (1997) base their models on the believe "that EQ is a set of skills that are learnable", which identifies their models "as trait models" (Ackeley, 2016, p. 271).

All models define a list of up to 18 skills or abilities required for EQ, divided into four to five categories, while there is still a huge overlap among the skills across the three models. The abilities across all models always address emotions of oneself, emotions of others and represented emotions, e.g., by objects like pictures (Ackeley, 2016). All models involve skills required to identify and know or understand emotions, manipulate, regulate, and manage emotions and skills regarding interpersonal or social relationship management. Within all models, mediation of emotions or through emotions is involved.

Logic and emotions seem to be inseparable according to Ackeley (2016, p. 271). She sees EQ as “the integration of emotions with cognition to impact performance”. Meaning is defined based on what is important. What is important is depended on emotion. Hence, EQ can be considered as the intelligent use of emotions. However, she concludes that every model has its validation, but none of them can capture the whole concept of EQ properly. One might need to merge all the models into one to capture it best. Birks & Watt (2007, p. 368) came up with a definition of EQ that might well capture the most important aspects, since it captures a mixture of all the skills EQ is proposed to require:

“[Emotional Intelligence is] a set of abilities (verbal and non-verbal) that enable a person to generate, recognize, express, understand and evaluate their own and others’ emotions in order to guide thinking and action and successfully cope with environmental demands and pressures.”

Still, this definition is superficial and is not able to capture all the nuances and facets the concept of EQ entails. For example, the term “generate” might be a bit misleading, since one can generate emotions within oneself, but one cannot generate them within others. Although one might be able to intentionally cause a specific emotion within others, by showing behaviour that might trigger this emotion to emerge.

All models aim at making EQ measurable, comparable to an intelligence test (IQ test), they aim at measuring emotional intelligence. They want to measure how quickly and effectively people understand, interpret, process, and make use of emotional information (Abbou, 2023, p. 65). The new promise of success via EQ led to capitalisation of emotions. Companies came up with concepts to make profit out of emotions and teaching the skill of EQ. The main field of application for EQ was business, where it was promising for success and effective leadership (Ackeley, 2016). The majority of literature that was found focussed on this field of application. However, EQ became more popular and figured out to be a useful skill in all kinds of fields. Birks and Watt (2007), for example, identified its importance and increasing value for patient-centred care within the medical sector. Since a lot of concepts and paradigms, as well as paradigm shifts within medicine are also found within psychiatry, it is more than likely that EQ also plays a highly important role within this field, especially, since psychiatry explicitly needs to deal with emotions of the patients (Frajo-Apor et al., 2015; Powell et al., 2015).

4.3 Artificial Emotional Intelligence

Now that I have elaborated on what human emotions are and how they arise and can be measured, I want to link them to the main topic artificial emotional intelligence (AEI). Therefore, I will first shortly explain artificial intelligence and then link it to emotionality.

4.3.1 Artificial Intelligence

Artificial Intelligence (AI) is a hard- or software with cognitive abilities, especially learning and problem solving. Math and statistics combined builds an algorithm, an instruction for the machine, expressed in the format of either code, tone or flow chart equip AI with those abilities. AI is mainly used to generate knowledge out of huge amounts of information. It is able to do so, because it uses different machine learning strategies, ranging from supervised learning to deep learning and thereby improves and adapts its action patterns (Abbou, 2023, p. 19-25; Cohen, 2021). AI helps in collecting, modelling and analysing data, among others via clustering to assist humans in e.g., decision making processes. To do so, AI computes probabilities of outcomes and produces predictions and adapts patterns according to new data input. AI tests and simulates all possible scenarios until it finds out the best possible solution. To be able to conduct such cognitive tasks, AI was designed according to human example. The processes happening within the system should simulate human cognitive functioning. Often it is said that AI is a black boxed technology, since people cannot understand what is happening, how an outcome was computed or which input or training data was used (von Eschenbach, 2021, p. 1608). Consequently, people do not trust the technology and are critical and hesitant regarding its wider use and application. So, how is it possible for a machine, which works with fully rule-based logic to be emotionally intelligent?

4.3.2 AI and Emotions

As described before, Ekmann's classification made emotions measurable, which means that emotions are transferred into data. Data is the basis of AI; hence, these data can be used as input for AI systems to open up the possibility for them to acquire the cognitive skill of understanding human emotions. This field of research is called *affective computing* or *Emotion AI*, which currently receives a lot of funding for research, so that it is rapidly developing and making progress (Abbou, 2023, p. 70). However, this field was already established in the 1990s, around thirty years after Ekmann introduced his approach. Rosalind Picard was a pioneer within this field. Together with Rana el Kaliouby, she was able to implement and automate the *Facial Action Coding System (FACS)*, which was developed by Ekmann and Friesen into computers (Baltrusaitis et al., 2011; McDuff et al., 2010). The FACS is a system to recognize emotions based on mimic musculature. Every muscle movement or combination of multiple muscle movements is an *Action Unit* (AU). The analyses of facial expressions resulted in 44 AUs that can be used to identify emotions on five stages of intensity. Further, feigned emotions can be identified, because the face needs time to pretend and the real emotion is expressed first for a second. This is a micromimicry or microexpression and can be measured, by a machine better than by a human. So, in the end the seven basic emotions can be identified with the FACS and Picard and el Kaliouby were successful in automating this process and even identify the real emotion behind microexpressions, which can also be considered as regulated emotions. However, the success of the system might also be tied to the laboratory settings (Misselhorn, 2021, p. 24). Pictures had to present a person frontal and the quality needed to be very high, so the picture is sharp. In real-world settings, the system was not that accurate. Another source to train AI in emotion skills is Big Data Training (Misselhorn, 2021, p. 23)

Despite emotion recognition via facial cues, AI is also able to identify emotions via analysis of speech patterns, sound analysis of spoken and non-spoken words, tracking of eye-movements and measuring neurological immersion (Abbou, 2023, p. 71). And thereby at least acquired the EQ skill of identifying and (partly) understanding emotions in others (Ackeley, 2016, p. 272-274). While face-based emotion recognition is limited in its area of recognition, because it is dependent on the reference object as well as social and cultural contexts, other ways of emotion recognition bring different sets of beneficial skills with them. Voice-based emotion analysis for example is promised to allow insight into the intimate central aspects of inner workings of personality (Misselhorn, 2021, p. 25). Sentiment analysis is aimed at analysing language to figure out the meaning of spoken claims. It can also be used to manipulate and generate emotions (ibid., p.29; 32). In general, the diverse technologies that might be a combination of several subsystems (multimodal systems), are not limited to one application area, but spread into diverse directions of human life.

Regarding the EQ skills of self-expression and self-awareness, Kenza Ait Si Abbou has a strong position. It does not play a role if a machine itself feels, as long as it makes humans believe that it feels (Abbou, 2023, p. 69). She thinks that the current technical state of machines is far from machines having consciousness. Machines would need to reflect about that they feel to make them conscious about their feelings. This ability to be conscious about the self is exclusive to humans and Abbou doubts that machines will ever reach it (2023, p. 68). Machines work based on simulations and computations of probabilities, hence they will only be able to simulate human emotions, probably very convincing, but they will not be able to genuinely feel the emotions themselves, since they will not reach the self. Schuller and Schuller have a different opinion about that. They work on emotion recognition, but also on emotion generation and talk about AI having “inner” and “outer” emotion (Schuller & Schuller, 2018, p. 40). This would provoke the conflict of the expression of “real” emotions and enable the AI to supress emotions, to hide them and let the humans believe they are in a different emotional state. This poses some ethical problems since the AI is manipulating humans and that affects humans and changes something within humans’ emotional states. Technology is not neutral and Artificial Emotional Intelligence would be even more powerful than every technology that is currently out there, since it can reach, affect, and manipulate humans’ most vulnerable parts, their emotions.

Moreover, as has been shown, emotion expression and perception seem to be bound to a body. Facial expressions produce a visual cue, verbal expression requires the ability to talk, and the root of emotions lies within neuron activation and hormone release, which’s effects can be perceived within the body as emotions. An AI not necessarily has a body or the ability to express verbally, consequently some requirements for emotion expression or emotion perception might not be given. Unembodied AI is likely to only have limited EQ skills or lack the ability to acquire some EQ skills as self-expression at all. Additionally, emotion identification in others is already difficult enough for humans, because of the subjectivity and context dependency of emotions, which makes it even more difficult for AEI to identify emotions correctly.

However, current developments of AEI mainly focus on and are limited to emotion recognition and partly involve reactions to human emotions in the context of conversational agents (Schuller & Schuller, 2023, p. 41). Most of AEI applications are currently unknown to the public. To get an overview of what can be assumed to be possible in practice, expert interviews were conducted, which will be presented in the next section.

5. Problem-Centered Expert Interviews

5.1 Method of Expert Interviews

To ensure practical validity of the research, problem-centered expert interviews were conducted. The qualitative method of conducting expert interviews has proven to be effective in exploring specific fields of action. Usually, expert interviews are merely theory generating and discuss the social relevance of expert knowledge (Döringer, 2020, p. 265). Problem-centered expert interviews in the contrary go beyond this and combine the classic approach of expert interviews with the methods of problem-centered interviews. This allows for a more discursive and dialogic interview procedure, in which personal perspectives of the individual experts can be investigated (Döringer, 2020, p. 265). This approach is especially suiting for this research because the field of AI in psychiatry is still rather visionary and currently is still lacking actual examples of application. Hence an approach allowing for more discursive components promises deeper insights and more flexibility. Due to the same reason, it was decided to follow a semi-structured interview guide, so flexibility is ensured, and the direction of the interview can be adapted accordingly with the experts' specific field of knowledge (Wilson, 2014).

All five experts were selected, because they work as professionals or researchers in related fields to psychiatry, psychotherapy, emotions, and AI. Experts were approached via mail after identifying them as potential interview partners via internet search or via snowball-sampling in the networks of experts already being interviewed. Two interviewees are male and three are female. Three interviewees are researchers and two work as practitioners in psychiatry and psychotherapy. Four interviews were conducted in English with experts based in the Netherlands and one was conducted in German with an expert based in Germany. The aim was to gain insights into the practical and ethical implications of Artificial (Emotional) Intelligence in psychiatry and psychotherapy from a practitioner's and researcher's perspective. The interviews took around 45 minutes and were recorded and transcribed afterwards. Unfortunately, for one interview, the audio recording quality was so poor that transcription was not possible. The analysis of this interview was based on the notes taken during the interview and very short fragments of the recording that were of little better quality. The transcripts of all other interviews as well as the interview guide can be found in the appendices.

5.2 Interview Results

The most important results of the expert interviews are shortly summarised in Table 1. More detailed results are presented individually per interview.

Table 1

Short summary of the interview results.

Interviewee	Prof. Dr. Dr. Kai Vogeley	Dr. Ed de Bruin	Caroline Bollen	Anna van Oosterzee	Marloes Veldhuis
Background/ Relation to psychiatry and AI	Clinical psychiatrist and neurologist, working with a broad spectrum of mental disorders, professor at Department of Psychiatry and Psychotherapy, link to technology in research	Psychologist, sleep researcher, assistant professor, teacher of compassionate technologies in the positive clinical psychology master's track	PhD candidate, research on empathy and communication technologies and neurodiversity	PhD candidate, research on AI in psychodiagnostics	Psychologist working broadly with all kinds of patients and mental illnesses, experience with e-health
Specific AI Applications	Fighting loneliness by using virtual friends, communication training for all kinds,	Chatbots, individualised online therapy, warning tools for e.g., narcolepsy, epilepsy	Large Language Models (ChatGPT), Voidpet, emergency aid technologies, meditation apps	AI as diagnostic tool, AI in psychotherapy,	NiceDay, Minddistrict
Chances/ Possibilities	Implementing diversity of therapy techniques in virtual avatars, almost every area that follow rule-based patterns, instructing meditation, conducting anamneses talks	More effective, faster, and individualised therapy and diagnosis (chatbots and online-therapy), creation of rich data and knowledgebase, supportive tools for therapist and patient	Aid in emergency situations, bridge way to enter therapy, support systems	Tools with predictive value, more nuanced, individualised therapy, emergency aid tools, creation of rich data and knowledgebase,	Easier access to therapy, potential to keep patients engaged in e-health, for milder conditions, monitoring tasks,
Challenges	High realism, lack of intuitiveness and flexibility of human	Human adoption of the technology, privacy, money and capitalisation,	Lack of consensus on definition of empathy, AI can never really	Diagnostic tools providing false reliability while lacking scientific	Keeping track of the progress of patients, emergency situations,

	therapist, transference love and attachment issues, over-reliance and addictive potential, lack of human interaction, keep human in the loop	structure of mental health care sector, conservativeness of therapists	substitute real human interactions, overreliance leads to prevention of development of important skills, power relationships, possible exploitation of vulnerabilities, commercialisation, privacy	foundation, overreliance on AI, privacy blocking research, oversimplification of concepts, classification, structure of healthcare system, emotion mimicry not necessary for effective therapy	question of responsibility, capitalisation, overreliance and setting boundaries
Desirability	Optimistic, especially since application cases of e.g., VR in anxiety treatment shows great success	Optimistic and enthusiastic, that there are a lot of possibilities, but still keep the human in the loop and do not let AI act independently	Tendency towards not desirable, too many pitfalls, opportunities should be explored, need for better health care options	Great opportunities for therapeutic tools and tools with predictive value	Open to learn about new possibilities if pitfalls are cancelled out
Attitude/ Perception of patients	Bias, because participation on voluntary basis	Willing to accept and use the online therapy options offered, motivational factor that *they want to be helped*	-	No practical examples, since the technology is not in practice yet	Willingness to engage with e-health low, increase in willingness expected when AI is implemented
Requirements for successful implementation	Keep the human in the loop, always provide supervision	Shift in field of psychology (away from traditional classification and conservative practice), technological maturity, give up something in exchange (privacy), keep humans involved	Keep human centered, focus on improving the health care system, integrate diversity	Change healthcare system, resolve privacy issues, only pursue tools with predictive value and scientific justification	Proper education of patients on self-responsibility to engage with application as well as possible risks, ability of AI to create warm connection, monitoring of applications by the government

5.2.1 Interview Prof. Dr. Dr. Kai Vogeley

Prof. Dr. Dr. Kai Vogeley, a clinical psychiatrist and neurologist at the University of Cologne's Department of Psychiatry and Psychotherapy, specializes in various mental disorders like schizophrenia, depression, suicidality, and autism. He has ventured into the scientific exploration of technology's role in psychiatry, particularly in using VR technologies and avatars to study joint attention, focusing on gaze behaviour reactions. These interactions were rudimentary but proved rewarding for participants, as evidenced by brain activity observed in fMRI scans. Emotionality is operationalized in his studies using virtual avatars to maintain experimental control, which would be impossible with human agents. Notably, participants found the interactions realistic. Vogeley acknowledges a potential bias in his study sample due to voluntary recruitment, suggesting participants were more open to technology. He and his team are also researching the social acceptance of new technologies.

Although his research is fundamentally scientific, Vogeley recognizes the considerable ethical and clinical challenges in applying technology in psychiatry. He sees AI's potential in guiding meditation, conducting patient interviews, and integrating diverse therapeutic approaches into virtual avatars, but only with supervision. AI could handle routine processes, combat loneliness, and provide support for expanding behavioural repertoires. However, the remarkable realism of AI-simulated virtual agents poses challenges, especially for psychologically ill individuals who may struggle to distinguish between real and simulated interactions.

Vogeley raises concerns about the addictive potential of AI applications and the lack of AI's intuition, flexibility, and adaptability compared to human therapists. AI may miss nuances and fail to detect the seriousness of implicit claims, necessitating human supervision to prevent crucial aspects from being overlooked. He also mentions the Freudian concept of transference love and warns against excessive dependence on AI, leading to a potential loss of competence in genuine human interaction. Despite these challenges, Vogeley remains cautiously optimistic about AI's use in psychiatry, particularly in anxiety disorder treatment, where VR has already proven valuable. VR has made strides in anxiety treatment and shows promise for broader psychiatric applications.

5.2.2 Interview Dr. Ed de Bruin

Dr. Ed de Bruin, a psychologist and assistant professor at the University of Twente's Department of Psychology, Health, and Technology, specializes in researching sleep interventions for children and adolescents. He is also a mindfulness trainer and contributes to courses on Compassionate Technology, addressing the use of technology in supporting mental health (de Bruin, 00:00:18). In his research, de Bruin explores how to measure well-being using technology and its physiological and psychological aspects. He conducted a study comparing online therapy's effectiveness for adolescents with sleep issues, finding that both face-to-face and internet-based therapy with automated exercises were almost equally effective (de Bruin, 00:41:30). This suggests a willingness among adolescents to accept online therapy, as they get the impression a real professional is concerned with their problems.

De Bruin mentions as possibilities for AI applications in psychotherapy, the use of chatbots. He notes that chatbots, with their history in psychotherapy, provide a responsive and helpful experience for patients (de Bruin, 00:03:19). He envisions chatbots serving as psychotherapy add-ons, offering psychoeducation and reflective support (de Bruin, 00:03:19). Moreover, he sees potential in AI-driven just-in-time adaptive interventions, which can tailor techniques to patients' needs (de Bruin, 00:03:19). De Bruin also emphasizes the potential of AI to extend public knowledge via AI-based medical and mental health databases (de Bruin, 00:21:49). Traditional methods of diagnosing mental disorders, practiced for over 40 years, face criticism about their validity (de Bruin, 00:22:07). AI could enable treatments that align more closely with patients' needs, and patients, motivated by a strong desire for help, are generally open to adopting new technologies (de Bruin, 00:28:34). He highlights a promising example related to narcolepsy, where temperature shifts before an attack could be used to integrate a warning function into wearable technology like a ring to warn patients before an attack is happening (de Bruin, 00:29:42). Similar applications might benefit epilepsy patients and others with related conditions.

Despite these opportunities, de Bruin acknowledges challenges. Therapists can be hesitant to embrace new technologies, because especially psychotherapists can be very conservative and fearing a loss of the human touch (de Bruin, 00:35:41). Additionally, slow transitions, financial complexities, and the need for justification in the mental health care field present obstacles (de Bruin, 00:13:14). However, he views these challenges positively, as they highlight the untapped potential in the domain and encourage societal action to bridge the gap between research and reality (de Bruin, 00:13:14).

Privacy concerns arise when AI systems process data on e.g., American servers without European data protection laws. De Bruin suggests that privacy may need to be sacrificed to enable revolutionary AI possibilities, considering it a concept of lesser importance (de Bruin, 00:25:38).

5.2.3 Interview Caroline Bollen

Caroline Bollen, a PhD candidate at Delft University of Technology, is researching empathy, communication technologies, and neurodiversity. Her work seeks to create a more inclusive conceptualization of empathy, particularly considering autistic empathic experiences (Bollen, 00:04:48). Bollen examines how communication technologies can either facilitate or hinder empathy (Bollen, 00:00:42). Using a virtue ethical approach, she views empathy as a virtue that can help address differences and similarities in experiences (Bollen, 00:05:24). Bollen emphasizes that the skills required for empathy are dynamic, influenced by technological changes (Bollen, 00:26:50). For instance, the shift from face-to-face to text-based communication demands different skills, such as interpreting emoticons.

Technological advancements are altering human interactions and our expectations of them. Chatbots, for example, offer endless patience, potentially leading to a shift in valuing technological traits over human characteristics (Bollen, 00:12:11). While Bollen believes that human-technology interactions can never replace genuine human relationships, they can change our expectations regarding technology's responsiveness (Bollen, 00:13:43). However, she warns that over-reliance on technology may hinder the development of empathy and lead to short-term solutions (Bollen, 00:15:42).

Bollen argues that technology should align with intentions to do good and reduce biases related to gender and ethnicity. She raises concerns about the narrow norms and exploitation of emotional technology markets targeting vulnerable individuals (Bollen, 00:09:32). The power dynamic between users and developers in terms of pricing and data collection, along with the parasocial nature of user-technology relationships, makes her hesitant to label these interactions as being empathic (Bollen, 00:24:23). The empathic components of the technology still come from the designers, not the technology itself.

Despite this, the mimicry of empathy can have practical benefits, especially in psychotherapeutic contexts (Bollen, 00:20:41). While AI cannot replace human therapists, it can help reduce the barriers to seeking therapy (Bollen, 00:09:32). In some cases, there may be less stigma associated with using technological tools for therapy than going to actual face-to-face therapy with a human therapist (Bollen, 00:31:56). However, Bollen emphasizes that technology should always play a secondary role to human support (Bollen, 00:16:40).

She suggests keeping humans at the centre of technological developments, using technology as an aid to address the strengths and weaknesses of being human (Bollen, 00:18:20). For example, during a panic attack, a technology aiding calming and soothing exercises could be beneficial (Bollen, 00:18:20). However, the technology not necessarily has to be intelligent, the aid can also be human recordings, like it is implemented in meditation apps like Headspace. The users still remain the intelligent part in this interaction and can choose themselves what suits them in that specific situation. While technology can assist, Bollen cautions against overreliance, as it can hinder the development of relevant problem-solving skills (Bollen, 00:24:23). She emphasizes that using AI and other technologies should not serve as a substitute for the lack of practitioners; each case should be assessed individually with a focus on improving the mental healthcare system overall (Bollen, 00:30:24).

5.2.4 Interview Anna van Oosterzee

Anna van Oosterzee, a PhD candidate at Utrecht University Ethics Institute and Leiden University in the ESDiT gravitation project, is researching how AI can aid in diagnosing mental disorders, particularly using supervised machine learning on brain scans (van Oosterzee, 00:02:35). Her interdisciplinary approach, combining psychology, neuroscience, and philosophy of psychiatry, examines the potential of AI in psychiatry. However, she notes that her diagnostic support technology is not yet in clinical practice and probably will never be (van Oosterzee, 00:02:35).

Van Oosterzee argues that developing diagnostic tools for mental disorders, as she is doing, may not be the right approach due to a lack of scientific foundation from fields like biology (van Oosterzee, 00:18:21). Unlike oncology, where AI automates diagnosing processes based on existing knowledge, psychiatry lacks established diagnostic brain patterns of mental illnesses for AI to identify (van Oosterzee, 00:16:04). She expresses concerns about over-reliance on AI diagnoses leading to false reliability and patients and psychiatrists getting a false sense of security in their judgement (van Oosterzee, 00:10:50). Van Oosterzee believes that a more effective use might have AI tools in psychiatry, which have predictive value, especially in therapeutic contexts (van Oosterzee, 00:07:55). She also questions the usefulness of rigid classification systems like the DSM in favour of more personalized treatment (van Oosterzee, 00:05:21). AI could help provide immediate support during specific moments, such as panic attacks (van Oosterzee, 00:07:55). Research shows that the most efficient things that can be offered at the moment are smaller tools that help with self-expression, self-development, and emotion management. These technologies could be made more accessible for people, who do not get into therapy, because of the complexity of the healthcare system (van Oosterzee, 00:36:07). However, capitalization of the healthcare system remains a problem. People in power over the money-flow might push expensive tools over effective ones. Consequently, the government and healthcare system should feel responsible to protect *true* therapy.

She stresses that AI should not replace human interactions but should supplement them (van Oosterzee, 00:38:58). In the mental healthcare system, combining AI technology with in-person therapy can potentially provide more sessions to patients (van Oosterzee, 00:38:58). The adoption of these approaches depends on trial and error, since the success of technological intervention is never guaranteed (van Oosterzee, 00:46:58). Van Oosterzee sees the potential for collecting longitudinal data through wearables and phones but acknowledges concerns about biases, privacy, and data security (van Oosterzee, 00:07:55). She thinks that privacy laws are often blocking research while being not the most important ethical concern. Therefore, she suggests that privacy laws should be balanced with the priority of tools to prevent suicide and other critical mental healthcare needs (van Oosterzee, 00:22:45).

Regarding emotional and empathic components in psychiatric AI tools, van Oosterzee believes these may lead to uncanny valley issues and are not necessary for effective therapy (van Oosterzee, 00:29:02). Therapists are meant to provide guidance without entering into the patient's emotional state, the same will be the case for AI (van Oosterzee, 00:31:41). While some level of sympathy is important, emotion mimicry is not (van Oosterzee, 00:33:06). Further, emotions and empathy are very complex concepts, and each discipline has different theories about their nature. Oversimplified versions of those complex problems need to be created to translate the complexity of theories between the disciplines and van Oosterzee is worried that society is relying only on the simplified versions of the complex concepts, which limits the dynamic development of knowledge (van Oosterzee, 00:34:40).

5.2.5 Interview Marloes Veldhuis

Marloes Veldhuis, a master's psychologist with five years of experience, has been working at MensGGZ for three years now, providing psychological care for a broad range of patients without specializing in one area (Veldhuis, 00:13:50). She is also experienced in using e-health technologies, primarily NiceDay and Minddistrict, though neither fully satisfies her

(Veldhuis, 00:13:50). NiceDay, while offering fully online treatment, focuses too much on cognitive behavioural therapy (CBT), while Veldhuis prefers more diverse approaches like acceptance and commitment therapy (Veldhuis, 00:13:50). Minddistrict is not as flexible as NiceDay but has more material to offer. It is more structured, and program based, e.g., there are programs for anxiety or acceptance and commitment. However, it was to cost intensive, and their company stopped using it (Veldhuis, 00:19:23).

One of the major issues in implementing e-health programs is patient engagement (Veldhuis, 00:15:46). Patients must understand their responsibility for following e-health programs, but Veldhuis observes that when patients are still in contact with a psychologist, they engage less with e-health tools (Veldhuis, 00:17:20). She sees potential for AI to recognize patterns of patient disengagement and intervene (Veldhuis, 00:08:21). Veldhuis highlights the importance of establishing a warm connection between patients and therapist, which e-health technologies currently are lacking (Veldhuis, 00:17:56). She believes AI can enhance this connection and make interactions more interesting and rewarding (Veldhuis, 00:17:56). Currently, e-health products lack rewards or reciprocity, making them less engaging (Veldhuis, 00:33:47).

However, Veldhuis is concerned that overreliance on e-health systems may hinder patients from becoming independent again (Veldhuis, 00:33:47). In a human therapeutic relationship, the human therapist decides when to end, while companies providing these AI applications may not encourage users to quit, because they are making profit out of it (Veldhuis, 00:35:11). The issue of responsibility in e-health products is complex, particularly in critical situations (Veldhuis, 00:08:21). It is uncertain how the provider of the system should react and be held responsible. For example, when a person is texting "I am not doing well, I am on top of a bridge and want to end things." How would the providing company step in? This question remains unresolved currently and responsibility cannot clearly be assigned.

Veldhuis notes that AI is better suited for milder conditions where patients have more resilience (Veldhuis, 00:10:51). For example, AI can be used for long-term mood monitoring in patients with bipolar disorder (Veldhuis, 00:11:55). The success of e-health depends on patient engagement and the specific application (Veldhuis, 00:19:23). Education is crucial for patients to safely use e-health applications, and government monitoring can ensure the systems' validity and safety (Veldhuis, 00:38:19). AI has the potential to bridge the gap in the shortage of professionals and make therapy more approachable (Veldhuis, 00:04:20). Veldhuis envisions AI as a valuable support tool for therapists but wants to maintain control over it (Veldhuis, 00:35:50). From a patient's perspective, user-friendly and accessible applications can make the process of seeking therapy easier, especially in cultures open to new technologies (Veldhuis, 00:35:50).

5.3. Analysis of the Expert-Interviews

The expert interviews provided rich insights into the possible practical and ethical implications of applying (emotional) AI in the psychiatric context. All experts agreed on the possibility that AI might enable for more diverse and individualised approaches, which can be tailored especially on the needs of the patient, which is a promising factor for more effective and successful therapy. Further, almost all experts agreed on the opportunity that AI has the potential to be the bridge for easier accessibility of therapy and hence improving the situation of shortage of professional help. AI might help to overcome the stigma and taboo that is still often related to psychotherapy. Hence the experts believe that it might be easier and more acceptable for possible patients to use technology to enter therapy. Since people are used to their smartphones, they might be more open to give therapy a first try when making their first experiences via something they are familiar with (likely their smartphones with an AI-based app).

There also were some points, where the experts had divergent opinions. The creation of rich data- and knowledgebases was evaluated as a huge benefit by de Bruin and van Oosterzee, while other experts perceived the accessibility and collection of data as intimidating privacy. Both de Bruin and van Oosterzee do not perceive privacy as an important issue and claim to rather dedicate more importance to potential life-saving technologies than to protecting privacy. Values need to be weighed against each other and both believe, privacy would then be considered as less relevant than, e.g., saving lives. “I think that technology ultimately it will be of a great help. […] [W]e might need to give something up in exchange for that. I don't know what yet but maybe privacy […] is a concept that is not that important actually” (de Bruin, 00:25:38). Both agree that the collection of new types of data opens up new possibilities for knowledge creation and the scientific community to catch up upon and develop better solutions and support for patients based on the new data.

“I like the prospect of collecting a lot of data. Now with the wearables and your iPhone, you can really collect these fine-grained longitudinal data, which before was not possible. Like I remember when I started my study of psychology, there was all this complaining of, ‘Oh, you can't collect the type of data we need’. And now ten years later that's just […] not the case anymore. You can collect the type of data you need, and you can process this big data sets. It's amazing how fast this is developing.” (Van Oosterzee, 00:07:55).

Other experts, as Caroline Bollen are more concerned about the collection of sensitive data. This especially links to the position of power the people, who provide the technology, are in, since they decide "what they do with the data they gather" (Bollen, 00:24:23). This might imply that the providers exploit the vulnerability of the patients, who are possibly dependent on using the technologies for their mental health. Hence providers might tend to value their own profit over the safety and quality of their products. Patients might be unaware of this and blindly trust the technologies, which can do great harm. Overall, all experts agreed that capitalisation does play a role and might negatively influence the application of AI tools in psychiatry.

A point of disagreement between the experts was the application of AI tools for emergency situations and immediate assistance, as well as just-in-time adaptive interventions. All experts thought AI would be especially helpful in emergency situations to e.g., help patients calm down during panic attacks or receiving a warning before a narcolepsy attack is happening. Only Veldhuis raised concerns about using e-health in emergency situations. She is worried that the application or affiliated company is not able to sufficiently provide support and approach the patient in time, which also raised the ethical question of responsibility, which remains unresolved until now. "Who are you going to contact? Because you might not have any information at all, and you might not even have the contact information of their general practitioner. I think that in that sense, it might be dangerous." (Veldhuis, 00:29:06)

Another point of debate was the implementation or rather the necessity of implementing emotional intelligence into AI in psychiatry. Bollen was hesitant to say that the technology itself will be empathic. She claims that the impression of empathic interaction is created by the humans behind the technology, involved in development progress. De Bruin (00:26:25) similarly argued that "you might say that there is a residue of a human in the response of AI because it's all based on previous human. You know, generated information. […] there is still humanity in the response of AI, even though it's artificial intelligence."

Some experts saw great opportunities for AI helping to deal with emotion management and creating warm connecting trust relationships between patients and the technologies, while van Oosterzee perceived it as rather hindering to implement emotions into therapeutical applications. In her opinion, therapists are not meant to show too much of empathy in patient interactions and the involvement of emotions will rather hinder the success of therapy.

"So maybe an AI doesn't have to mimic that much emotions. If it wants to be an effective therapist, it just has to mirror what you are doing and to help your process through your options and your biases. But in a way, a therapist is also a bit empty. They're not really supposed to get their own emotions involved into the mix." (van Oosterzee, 00:27:05)

Besides the disagreement with some of the other experts, who see empathy as an important factor for success in therapy, emphasise the importance of empathy in therapeutical interactions and criticise that technology is lacking it currently. Veldhuis for example argues that current e-health technologies still lack the ability to connect with the patient in a warm, empathic manner and that, if AI would be able to establish this connection, it would help e-psychotherapy to be more effective and successful and get patients to engage more in using it (Veldhuis, 00:17:56).

It was agreed among some of the experts that the impression of human-involvement will be enough to help patients accept the technology. Both de Bruin and Bollen argue that the technologies used in psychotherapy not necessarily need to be intelligent for treatment to be successful. Bollen (00:19:33) says, that there are still the users, who are intelligent and can decide for themselves what is good for them.

All experts see an addictive potential within these kind of emotional AI tools in psychiatry. Vogeley for example argues that gaming addiction is anchored as mental disorder, meanwhile similar conditions might occur for AI technologies. Further, he refers to transference love and sees the related relationship regulation as a possible problem also with emotional AI technologies. Veldhuis similarly thinks, it might be hard for patients to cut off consulting the technology at some point, since this is already difficult with a human therapist. However, while a human therapist can decide, when to make the cut, AI might not, and users could become dependent and over-reliant on the technology. Also, the other experts see the potential problem of users' over-reliance on the AI tool in psychiatry.

Both Bollen and Vogeley are worried about a potential loss of important skills and capabilities of the users when they rely too heavily on AI interactions. Bollen (00:13:43) says that it might prevent people from developing meaningful human relationships, because they are not able to develop the skills needed for real-world human interactions. Vogeley similarly refers to the potential lack of real-human interactions. Nevertheless, he sees potential for AI to solve feelings of loneliness via virtual friends or avatars, who serve as company for the patients. However, Vogeley's vision goes beyond current applications as 'Replica' and aims at more controlled, safe alternatives, possibly supported by insurance companies.

In general, the poor infrastructure of the mental health care system was seen as a factor that hinders patients to receive help currently and opens opportunities for AI to bridge this gap of entering mental health care. Additionally, de Bruin and van Oosterzee also reported on research that is being conducted in the field, which brings about applications that will never make it into practice among others due to insufficiency of scientific foundation. Further, the lack of consensus on what the underlying concepts of what emotions really are was mentioned by all experts and is also due to the impossibility to find a proper scientifically based measurement. This does not only account for emotions, but also for empathy as Bollen reported, as well as for well-being as de Bruin reported. Moreover, the practiced classification of mental illnesses was perceived as insufficient by most of the experts since it has been harshly criticised and pursues stigma. De Bruin thinks that AI might help to tailor the diagnosis and treatment better to patients.

> “It might be much more applicable, it might even be a classification that is somehow adaptive to cultures, might be adaptive to age, to gender, […], to sex. It might be adaptive to individuals even and developmental issues. And then you can come up […], with diagnoses and treatments which are much closer to what a person, an individual needs.” (de Bruin, 00:22:07)

While van Oosterzee believes, that sticking to classification will not resolve the issue. She does not see potential in diagnostic AI tools, because of ethical concerns and false reliability, but visions tools with predictive value to be impactful in the field of psychiatry.

> “I've been mainly criticising the DSM system and I think […] there's a lot of possibility with these technologies if you let go of the classification, depression, or anxiety. I think that if we develop these wonderful like emotion recognising or mimicry systems, but then still try to diagnose depression that's just not going to add a whole lot more […]. Like you cannot move once you get into that corner, you cannot do anything with the classification depression. You're completely stuck once you adopt it.” (van Oosterzee, 00:06:10)

Overall, the willingness of patients to adopt AI technologies was rated as being likely, since they are open to try new possibilities of receiving help. However, all experts agreed on the necessity of the human-in-the-loop condition. So, there should always be a human behind the technology, taking control and responsibility if things develop in the wrong direction. Hence, AI in psychiatry should rather have the character of a supporting technology that supplements human therapy and should not be a fully automated agent. An argument for this was provided by Vogeley, who says that AI might miss the hidden message behind patient’s words. He thinks AI might perform well in basic tasks that follow routines. Veldhuis argues in a similar direction. She thinks AI might be useful in dealing with milder conditions and patients, who still have resilience, while severe conditions as suicidality require human lead. Hence, it can be concluded that emotionally loaded situations and tasks should rather remain handled by humans and emotional AI agents should not intervene. Veldhuis especially stresses the question of responsibility for cases that develop negatively due to AI intervention, which is still unsolved. In the end success of AI-mediated therapy is not guaranteed since every individual responds to and perceives it differently. (Veldhuis, 00:19:23) However, patients are "usually […] more open [to use technology for therapeutic purposes] […] they have one more motivating factor and that is they want to be helped."

Because of the worrying unresolved issues, some experts suggest that the government should take responsibility and regulate the use of AI in psychiatry to prevent further pitfalls as capitalisation of these technologies until issues as biases, lack of scientific evidence, possible overreliance on the AI and privacy are resolved.

6. Implications of AEI in Psychiatry

Despite all the promised possibilities envisioned for Artificial Emotional Intelligence, there are still a lot of challenges to be faced. The next sections will identify ethical and practical implications AI and AEI might have in psychiatry.

6.1 Ethical Considerations

Katirai (2023) recently reviewed ethical issues related to emotion recognition technologies (ERT), which already implies that systems, like the FACS, might not be as glorious as they are presented by their developers. She presents a body of literature, which is slowly emerging that raises concerns about the use of ERTs (Katirai, 2023, p.1). She describes that a shift is happening away from the initial application sector of healthcare to more commercialised applications. This might explain, why AI tools are scarce in the psychiatric sector currently and little to none are actually in use, while more and more AI-based applications for the more commercialised well-being sector are introduced (Ray et al., 2022, p. 4). Still major investments are spent into the development and implementation of ERTs, despite the unclarity of its future, because the upcoming AI Act proposed by the EU might restrict ERTs and at least ban their implementation in predictive policing contexts (Katirai, 2023, p. 9).

In general, clinical governance is needed for AI-driven technologies in psychiatry (Lovejoy et al., 2019, p.2). While medical products have to adhere to strict regulations, medical apps do not. To ensure, safe use of these applications, patients need to be properly educated on the use conditions, as strictly as for medication intake, which Veldhuis already advised (00:38:19). In the US there were almost 50.000 mental health apps available in 2015. However, most of them have never reliably been validated or only been tested “in small-scale, short-term pilot studies” (Lovejoy et al., 2019, p.2). Two issues go along with this fact. Several mental health apps are not labelled as medical products, since they focus on lifestyle and self-management, as mentioned before. Second, the lack of validation leads to poor quality of information spread through the apps, which might have harmful consequences, e.g., in terms of inadequate advice. Therefore, in the UK labels were introduced, which point out apps with significant evidence of effectiveness and safety by the *NHS Digital Apps Library*.

Further, monetary investments continue, although the underlying premises of those technologies have undergone harsh criticism and a potential danger is posed by the development of ERTs in the direction beyond merely recognising emotions, but also predicting and finally controlling behaviour. Three key ethical issues of ERT could be identified, which are the risk of bias and unfair outcomes, “the sensitivity of emotion data and […] the risk of harm” Katirai (2023, p. 3).

The risk of bias is not new or unique to ERT, is it a common, frequently occurring issue within AI technologies. However, in the context of ERT, Katirai (2023, p. 5) highlights that bias and unfairness of these technologies is rooted in the lack of consent about what emotions really are and how to operationalise and measure them. As described in the critique on the common view of emotions, Katirai claims that there is no universal definition of emotions and that the basic-emotion approach implemented into the technologies is only able to assess a very limited number of emotions, which is not representative at all. Similar, to Barrett et al. (2019), Katirai (2023) warns about the mismatch of emotion expression and emotion perception, while stressing that this mismatch does not only apply for others perceiving our emotions, but also for self-report.

She also sees a danger in amplifying the implicit bias of the raters of the training data, which is later used to train the algorithm. Exacerbating existing human biases and values misalignments, which are implemented via the training data, is not exclusively and issue with ERTs but also within possible psychiatric AI applications (Brown et al., 2021). Further, cultural bias seems to be very prominent, since often white Western people are doing the rating. Hence, Katirai (2023, p. 5) claims that ERTs are often a product of the “Western, Educated, Industrialised, Rich and Democratic (WEIRD) nations” and therefore produce results, which are hardly representative of the real-world.

The same applies for psychiatric data, which is often claimed to be flawed and biased (WHO, 2023), as well as the psychiatric practice of diagnosis, where classification systems are still used, despite harsh criticism regarding their accuracy and usefulness. Within the field of psychiatry, a paradigm shift is happening, since the limits of current (tools for) categorisations and classifications of mental disorders fell for harsh criticism for a while now (Cooper, 2004). Psychiatrists themselves often disagree with the AI-diagnosis or criteria for diagnosis in general, they rarely rely on strict categorical diagnosis anymore, which sparks doubt about AI's success in diagnosis, when relying on out-dated methods and data (Brown, et al., 2021). Here AI would still pursue human bias instead of adding value, by using alternative approaches. Already, some of the experts like de Bruin and von Oosterzee showed concerns about this practice to be pursued.

Also, the inapplicability of AI technologies in psychiatry outside of the training setting, which is often due to implemented biases, is a major challenge. Results often cannot be transferred from laboratory settings to the real-world. More research is needed to tackle these issues sufficiently. However, the studies of AI-driven interventions are mainly conducted by their developers, who may value their personal monetary profit over the adequacy of their product, consequently, the benefits might be biased until third parties conduct the studies and validate the products in psychiatric contexts (Ray et al., 2022, p. 4).

Movements started to correct these biases and reestablish fairness within the data, but still diversity is lacking in the samples and data, especially regarding age, disabilities, and nationalities. Consequently, the two main areas of bias and unfairness, Katirai (2023) identified, are deficiency, which includes selection and sampling bias and contamination, which includes historical, behavioural, representational, and observer-based bias. The vagueness of the concept of emotions and the lack in consensus about it remains important, since one cannot build a suiting model for AI or ERT implementation, based on erroneous, misleading, and non-evidence-based assumptions.

Another ethical issue Katirai (2023, p. 6) discussed is the sensitivity of emotion data. The data gathered and used by ERTs is highly sensitive and gathered by invasive measures into persons privacy. Some people frame it as 'mental data' and there are even movements that want to establish new privacy rights to protect the mental property (Bostrom & Sandberg, 2011, p. 21). Emotion data gives intimate insights into "what it means to be human, including questions of identity, autonomy and freedom of thought" (Katirai, 2023, p.6). Every emotion data is unique, which makes the persons, the data is derived from, so called data subjects, vulnerable and prone to manipulation, when their data will be revealed. Some scientists warn that the methods to gather emotion data will become even more invasive in the light of current developments of such technologies, for example when multi-modal sensing is included. Katirai (2023) claims emotion to be inherently private and worthy of being protected. Health-related data is considered to be highly sensitive and emotion data should be treated in a similar way, although this is rarely the case.

Psychiatry is a highly sensitive field of application for AI, since it involves very vulnerable individuals and highly intimate data and information. Since there is a high risk of stigmatisation and discrimination steaming from mental health data, if data is being disclosed, this type of data needs to be considered as particularly sensitive (Lovejoy et al., 2019, p. 2). Patients of psychiatry are especially vulnerable, because of their mental health conditions and often they are dependent on receiving appropriate support. Bollen raised concerns about making profit out of peoples' vulnerabilities (00:24:34). Moreover, mental health patients, depending on their mental health condition, might lack the ability to provide consent, since some mental illnesses affect cognitive capacities (Lovejoy et al., 2019, p. 2). In these cases, it raises the question, if their consent remains valid while their condition gets worse.

Additionally, the capturing of emotions is problematic, because one tries to construct an objective phenomenon out of something that is inherently subjective and contextual. As van Oosterzee explained, there is a danger that information about the concepts gets lost or oversimplified during translation across disciplines (00:26:02). Hence, the question is who should have access and control over the data? It is proposed to give the data subjects themselves the control over their data, also to circumvent the possible problem of emotion surveillance. Surveillance of emotions could have severe consequences, such as the loss of autonomy and authenticity, the reinforcement of emotion stereotypes, a possible "alienation" from the own emotions and an increase in social pressure for emotion regulation. The sensitivity of emotion data is even more of an ethical concern for vulnerable groups, which were already discussed by Barrett et al. (2019). Katirai (2023, p. 6) also hints at "inalienable 'affective rights'" which every individual has according to the United Nations Universal Declaration of Human Rights. Continuing using and developing ERTs in a way it is done currently, would intimidate the ethical design practice and ignore these widely recognized rights.

Another ethical issue identified by Katirai (2023, p. 7) is the use of ERTs in consequential settings. This involves field and sectors of application, where the results of ERT have real consequences for either human decision-making or the data subject themselves. Here the vulnerability of individuals again plays a key role, since people who want to profit from ERT in consequential applications often are seeking for help to solve a certain problem. This is for example the case in the healthcare sector, where ERT was initially the key sector of implementation before it expanded to more commercialised implementations. The application of ERT in this domain may threaten the autonomy of patients on the one hand, because the results might reveal thoughts and conditions the patient did not intend to share with the health care professionals. On the other hand, there is no clarity about accountability for possible misdiagnosis or non-detected diagnosis by the ERT. Here the question of responsibility already discussed by Veldhuis comes into play again (00:35:11). A rethinking of the concept of confidentiality of data from health context is required and an expansion of the data protection law for health data to account for the implications the use of ERT might have.

To tackle these ethical issues responsibly, Katirai (2023) suggests to first assess the ability and quality of the system realistically, as well as mapping potential risks. There is a need for an understanding of emotion expression that nuanced and acknowledges emotion expression as just one observable component of a more complex system. Hence, it should not be used for prediction purposes until consensus about the nature of emotion is found. This regulation of ERT is also a challenge in itself, especially on a global level. There is a need for oversight, the so-called human-in-the-loop, but many countries still need to establish safety guidelines regarding emotion recognition. The European Union tries to tackle parts of the problems within the European General Data Protection Regulation (GDPR), but emotion tracking is not directly included, while the upcoming AI Act as mentioned possibly will restrict biometric surveillance in general and certain forms of ERT specifically (Katirai, 2023, p. 8).

Discussions about the human-in-the-loop rule or even more concrete, whether AI-systems should be allowed to work autonomous and take over human tasks at some point are also prominent in the psychiatric context. The experts already addressed the issue in the interviews. Therapeutic apps, especially chatbots are the most popular domain within AI-based psychotherapeutic applications (Lovejoy et al., 2019, p. 2). These apps mainly mimic human behaviour in therapy, help the user to explore their emotions and mental conditions. Furthermore, they offer advice and can refer patients to psychiatric services close by when needed. Usually, AI-applications are used within human-supervised therapy. However, Lovejoy et al. (2019, p. 3) and colleagues predict that they will surely be applied more autonomously, outside settings, which underlie human supervision, although this will take time and is a slow process. Before this step will happen, aspects of ‘traditional’ therapy need to be re-thought and adapted to the new circumstances and possibilities (Lovejoy et al., 2019, p. 3). Since reports have been published about suicides and other harms, which were conducted after receiving therapy from generative language models, it seems to be more advisable to follow non generative models for therapeutic interactions, which most commonly known therapeutic chatbots, e.g., Wysa and Woebot do (WHO, 2023; Hale, 2023). Since generative models produce a new output with every interaction, clinical safety and validity hardly can be tested (WHO, 2023). Hence, among others, the Wysa chatbot uses statements, which were pre-approved or created by human therapists and additionally ensures patient's privacy and anonymity by refusing to collect data which could be used to identify the patient (WHO, 2023).

André (2023) warns that e.g., trauma patients as war veterans suffering from *Posttraumatic Stress Disorder* (PTSD), should not be treated by AI systems. She explains that it is unethical to expose those highly vulnerable persons to machines, since suffering should be seen and perceived by real humans. At certain points machines should simply not replace human agents. This seems to be an ambiguous statement since it is difficult to set a threshold. Especially suffering, as well as all other emotions, is perceived highly individual and can and should not be weight against each other (Hofmann, 2015). However, some of the experts, as Vogeley and Veldhuis, also advised to only use AI-based therapy for patients suffering from milder conditions, who still have sufficient resilience to tackle their problems themselves with the aid of AI-tools. Consequently, one cannot generalize where to not use artificial agents and where to use human agents. It might be a solution to keep the decision with the client to decide which agent they prefer. Besides, AI systems used in psychotherapeutic contexts are always supervised, so a human agent always has some control to intervene if necessary (Gebhard, 2023). AI systems thereby always only have an assisting function in psychiatry. Further, AI systems are developed based on statistical approximations, which means there is no guarantee for success of their use.

A positive effect of new technologies is not necessarily guaranteed since patients might perceive technological intervention differently. Veldhuis also scratched this problem and discussed that it is highly depending on the technology and the willingness of the patients to use the technology and whether it fits the patient (00:19:23). While some patients might feel empowered by the control, they gain via monitoring their illness, some patients might perceive it as overwhelming to have this additional responsibility and feel as if they are constantly reminded of their mental health condition (Lovejoy et al., 2019, p. 2). Another problem that might arise is possible over-attachment of patients to AI applications on the long run (Ray et al., 2022, p.4). This has also been experienced by the company clare&me, who provide therapy via voice-based chatbot. Due to the 24/7 accessibility of the service of the voicebot, users started to consult the voicebot for every decision and problem they had to deal with, which restricted them in their autonomy and made them dependent (Abbou, 2023). Further, it has been shown that patients treat assistive tools in a ruder way than they treat humans doing the same tasks (Ray et al., 2022, p.4). This might indicate less respect and acceptance for AI-driven therapy, which could also affect the success negatively.

In the review by Katirai (2023) it is also stressed that all the issues of insufficient scientific evidence, as well as missing consensus about the nature of emotion, acceptability and awareness of possible users for ERT, biases and power dynamics resulting of the WEIRD nature of the test data, need to be sufficiently solved, else ERT will not genuinely be helpful, but be just another quick technological fix for real-world problems that does not help on the long run.

While the *Regional digital health action plan for the WHO Europe Region 2023-2030* pointed out a need for innovation in predictive analytics for improving health by the means of big data and AI, WHO stresses that the current status of applications and research in AI for the mental health domain needs to be assessed (2023). To enable successful implementation, mental health care systems need to adapt their procedures and structures to enable advancement in mental health services. Although, "AI has the potential to significantly transform the technical aspects of psychiatry by discovering hidden patterns" (Brown et al., 2021, p. 132), it does not perform well for large, highly heterogenous samples and coming up with accurate predictions for uncommon and unfamiliar patterns is very difficult for AI. Humans can broaden their view and contextualize the information, they see the environment and background of the client, while AI has to work with the data given (André, 2023; Brown et al., 2021). AI most of the time works strictly data gathering, especially in a psychodiagnostics use context, which is a task that is usually done by the less-trained psychiatrists anyways (Brown et al., 2021). Hence, it is questionable, whether the application of AI for diagnostic tasks is suitable.

Moreover, if all these ethical issues remain unsolved and research, development and implementation continue, pretending to follow certain guidelines, "ethics washing" might happen, so companies claim their ERTs to be ethical to the public, without actually making any effort (Katirai, 2023). This would reinforce the profit-driven approach that just follows financial interests instead of using the ERTs to actually contribute good and help people. The interests of stakeholders need to be aligned with any potential user groups, especially vulnerable groups and their needs and East and West values need to be combined, to ensure fairness of the systems.

The harsh conclusion of some of the reviewed authors is to ban ERTs until the ethical issues are solved (Katirai, 2023, p. 8)

6.2 Lack of Scientific Evidence and Transparency

Another implication that is almost always present with AI, is the Black Box problem. As explained before, people do not understand which data is used and how the outcomes are generated by AI systems. AI often works according to input-black boxed processing-output schemas and therefore the way the data is processed is unknown to most people and so complex that even knowledgeable people would hardly understand all these processes. It is often referred to this issue as explainability of AI or XAI. Bergstrom and West (2020, p. 43) claim that outcomes can be called bullshit, when they are based on biased input as well as outputs that clearly conflict with the intended purpose of the system. Harry Frankfurt coined the term bullshit and defined it as nonsense information people make up to persuade or impress others. Often this information is covered in stories or scientific jargon, which others do not understand. Bullshit information frequently is presented as being scientific claims that are difficult to fact-check, because they are generated within a black box and almost no one can understand how. However, to detect the source of bullshit it is often not required to open the black box and understand all processes. It might often be sufficient to check whether the data used appropriately fits the problem and is carefully selected. Further, it helps to assess if the model and the purpose of the system are reasonable and unbiased (Bergstrom and West 2020, p. 43). Shortly, bullshit outcomes can be defined as nonsense results or claims that are based on non-existing or misinterpreted evidence. Most of the time, flawed data is used to back up these claims.

The quality of training data for psychiatric AI systems is untested and invalid to a great extent and in general there still is too little data on patients and clinical performance available. A WHO study on AI-based psychiatric technologies revealed methodological and quality flaws in the research of mental health AI systems (WHO, 2023). The use of AI-driven tools is unbalanced in mental health research. There is a significant gap, because AI tools are mostly studied for depression, schizophrenia, and other psychotic disorders, while sufficient understanding of the use of AI tools for other mental illnesses is lacking (WHO, 2023).

Complex statistics and high-dimensional data are used for AI application in mental health, which might come along with inaccurate interpretations of results, biases and in case of inadequate handling, to over-optimism about AI's performance (WHO, 2023). Moreover, data is infrequently validated, and replicability of AI models is undermined by the lack of transparency. The underlying models often remain unknown to others since researchers keep them private and do not collaborate with other researchers. This is one of the factors that currently still prevent safe and practical implementation of AI in psychiatry (WHO, 2023). However, Abrams (2023) points out that the rapid progress of AI technologies and the linked capabilities of the systems may outweigh the need for full understanding of the internal workings of those systems in the future.

In the case of AEI, the problem of missing consensus about what emotions are and how to approach the research around them are concerning. As discussed before, there was also no reliable scientific evidence found for the effectiveness of ERTs. So, if one trains the AEI with the contaminated and insufficient data discussed by Barrett et al. (2019) and Katarai (2023), the outcomes most likely will be bullshit and say nothing about the real-world. Moreover, if the models the AEI is based on are constructed based on unsound premises, including the basic-emotion approach, which has been proven to only represent limited facets of emotion, as well as lacking diversity and subjectivity, while still being based on no universal definition of emotion, there is no guarantee that this model represents the intended real phenomena of emotion expression and emotion perception. Hence, all results the system can provide based on the model will be bullshit.

All these unclear connections and mismatches between missing evidence and flawed-data can be drawn back to a phenomenon Sullivan (2020, p.1) calls link-uncertainty. Link-uncertainty is the lack of scientific evidence for the link that should connect the real-world phenomenon with the model. This is a common problem that occurs with black-boxed algorithms. To reach true understanding and true outcomes requires overcoming link-uncertainty. However, it does not necessarily require opening the black box, but rather investigating deeper into the systems underlying model and how the real-world phenomena and the related affected population are simulated by the algorithm (Sullivan, 2020, p. 5). Especially, the countless hidden-layers within AI systems make it difficult to establish strong link-certainties, therefore, it is even more important to rely on strong scientific evidence for the model’s simulations of the real-world phenomenon. Sullivan explains that link-uncertainty my arise when the model should account for several incidents, since it is unclear to which real-world phenomena the model refers and connections become blurry. Additionally, several modelled real-world phenomena could reinforce issues as involving high stakes, maintaining harmful stereotypes, and thereby leading to greater marginalization, because they might be prone to inductive risk (Sullivan, 2020, p, 28).

This also applies especially to the case of AEI, since emotions are highly subjective and therefore, hard to reliably model. There is no universal definition of what they are and how to measure them and the scientific evidence for attempted measuring trials, as ERTs, is very weak and consequently the evidence for the data produced by these ERTs, which could have possibly been used to model the phenomenon of emotion, is not reliable and representative. Hence, modelling emotions accurately and reliably is almost not possible and consequently, a strong link-certainty cannot be established. Also, inductive risk is likely, since during the history of ERTs a generalisation of false assumptions of the basic-emotion approach already happened and AEI systems will be trained with data containing this unreliable data, which only represents a limited number of the six stereotypes of emotion expression, as well as limited (cultural) representativeness in general. The connections of links within ERTs might also be very blurry, since, as Barrett et al. (2019) stressed, facial expressions do not exclusively account for only one emotion word. A smile might be one expression of happiness, but not the only expression of happiness and can also occur on other occasions, while the person smiling is in a totally different emotional state. This might be difficult to model and hard for the system to process and accurately distinguish.

A challenge that AEI systems will face especially in psychiatric contexts is that all human emotions are regulated, which makes it very difficult for psychiatrists to understand them and even harder for AEI systems to identify those emotions correctly (Gebhard, 2023). It has been claimed that the FACS is able to identify microexpressions, which are supposed to be the real emotions behind the regulated emotion, accurately, so there might be possibilities to overcome this challenge (Abbou, 2023). In this case, AI would be a helpful tool to complement humans by cancelling out human shortcomings. However, the scientific premises the system is based on have undergone harsh criticism, as has been discussed earlier (Barrett et al., 2019). Hence, the system needs to be used with caution. Link-uncertainty can be extended here to the specific mental conditions and diagnosis of the patients, which might be ambiguous and not necessarily uniquely true and identifiable. Further, they might be prone to pursue stereotypes of mental illnesses when being pursued even in AI-based psychiatry.

To sum up, the detected lack of reliable scientific evidence of Barett et al. (2019) and the persistent lack of consensus about the nature of emotion, their perception and measurement might prevent accurate modelling of this phenomenon. This would consequently mean that any kind of model of emotions implemented into AI systems might be unreliable and inaccurate, which prevents it from producing accurate, true, and effective results. Nevertheless, research and developments in this domain are pursued ignoring the lacking scientific foundation. This needs to be seen as highly critical and ethically questionable. Consequently, further research on the establishment of a reliable body of scientific evidence about the nature of emotion and in a first step the correctness of emotion perception, since all other technological developments rely on ERT in the first place, is crucial to enable AEI. Biases and stereotypes need to vanish, by more extensive research on vulnerable and marginalized groups to include them in the training data as well and appropriately represent the target phenomena.

6.3 Lack of Empathy and Compassionate Relationships

Empathy is an important skill humans use during their social interactions. Similarly, to emotions, there is no universal definition of empathy. Bollen (2023, p. 6) reviewed 111 articles about empathy and found "31 meaningfully unique understandings of empathy". Many of them discussed that empathy might be either of affective or of cognitive nature. I will use the definition of Misselhorn (2009, p. 351), which defines empathy as the ability to put yourself into someone else's shoes emotionally. One evokes feelings of the other person within oneself. Misselhorn (2021, p. 46) distinguishes between the two natures of empathy also mentioned by Bollen (2023). First, there is cognitive empathy, which means that one can rationally understand the emotions of another person. Second, there is affective empathy that means, to feel the emotions jointly, which automated ERTs are not able to. These systems usually serve manipulative purposes. Emotional AI should get someone to act out a certain behaviour. According to Misselhorn (2021, p. 48) there are three necessary and sufficient conditions that all three need to be present to prove empathy to occur. These conditions are congruence, which means one experiences a unity with the emotions of another one, asymmetry, which means that you know that you are only perceiving the emotions, because someone else is perceiving them and other consciousness, which means that you are aware that you are currently perceiving the emotions of someone else.

Recently, a research line regarding Artificial Empathy has been established, which goes beyond the approach of common ERT, because it acknowledges the body as an important part of the establishment of empathy (Misselhorn, 2021, p. 62). Therefore, embodiment, either in forms of robots or virtual agents is necessary to ensure social interaction of the systems with their environment and individuals. Application areas of artificial empathy require high quality of social interaction because this ensures that the artificial system is accepted as social counterpart (Asada, 2015). These systems are based on computational models of two kinds, either based on a theory-based approach or on a data-driven approach.

The first approach is based on psychological and neurological theories of interhuman empathy. These are computationally modelled and implemented into an artificial system. The problem with this approach is that it relies on abstractions of different research teams. While transferring these into the system information loss is likely. Additionally, the implementation itself might shorten and distort the theories. Moreover, it cannot be neglected that the conceptualisation by different teams might lead to different conceptualisations and implementations of the phenomenon (Misselhorn, 2021, p. 65-67). Van Oosterzee (00:36:02) discussed concerns regarding the loss or simplification of theories by transferring them across disciplines or into systems as a critical point as well, as can be seen in the interview section of this paper.

The second approach is based on Data from empirical research about the empathic interaction of humans. This data is used to train the system to recognise usual patterns of empathic behaviour. Alternatively, data about the reactions of users to empathic artificial systems can be used as training data. This approach is very data-intensive and requires careful selection of the data used. Furthermore, the data is context specific, and it is difficult to transfer it to other situations.

Insights from the commercial sector show that for artificial systems it is not necessary to perceive or even simulate empathy themselves, it is more important that the systems recognise emotions correctly. This was also shown by studies de Bruin reported about. As long as patients have the impression there is a human behind the system, checking what they feed into the system, they are satisfied with the interaction. Although there might not be a human involved, but the system is simply parroting what the patient says or mimicking human responses (de Bruin, 00:03:19). The NICA robot, a social, empathic care robot does not make the claim to perceive emotions, the artificial system just simulates them. However, if the reaction of an artificial system to emotions is appropriate, it is sufficient for people to ascribe the ability of empathy to them, although artificial systems will ever lack the three necessary characteristics of congruence, asymmetry, and other consciousness to be empathically enabled by definition (Misselhorn, 2021, p. 79).

Pashevich (2021, p. 584) contrary argues that due to the deceptive nature of emotions expressed by robots, it is harder to establish affective empathy with them. The affective expressions of robots are not convincing enough, consequently, humans would not fall for the illusion and stop perceiving empathy for the robot after a short period of time if they were tricked by the robot into perceiving empathy. However, within treatment of patients experiencing mental trauma and conditions, it is crucial to approach them with human empathy and compassion, which AI-driven tools are lacking currently (Ray et al., 2022, p.4).

Elisabeth André (2023) sees AEI as a great opportunity for psychiatry. She claims that AEI could be used as an icebreaker in therapy. Emotions can be contagious, which means that people align to "behavioural synchrony" during social interactions by "showing a similar facial, vocal, or postural expression" (Herrando & Constantinides, 2021, p. 2). This does not exclude artificial emotions. Emotional behaviour of robots contributes to social bonding because people get engaged by the robots' emotions and adapt their own emotions accordingly. Further, André (2023) claims that emotions are required to see and analyse the bigger picture. Verbal utterings need to be considered, as well as the interplay with the nonverbal keys, which is where AI could assist. If an AI-driven tool will be able to successfully establish a trusting, compassionate relationship to the patient and convey the impression of showing empathy and understanding, it would be a huge step in the direction of successful AI-based therapy and treatment (Misselhorn, 2021, p. 78). Empathy and establishing a trust-relationship are considered as crucial factors for successful therapy. Some of the experts like Veldhuis agreed with this assumption, while van Oosterzee thinks therapeutical relationships should be more neutral and less emotionally loaden.

As can be seen, the views, regarding empathy and compassionate therapist-patient-relationships diverge and researchers are unsure about how users will perceive and accept the implementation of empathy into artificial systems. Reasons for this dissonance might be provided in the next section, where the uncanny valley problem will provide more insight into possible effects of perceived empathy with artificial objects.

6.4 Uncanny Valley Problem

Another implication that might arise from the use of Artificial Emotional Intelligence systems, is the problem of the uncanny valley. The problem of the uncanny valley was first described by Masahiro Mori in 1970. Ciano Aydin (2021, p. 300) explains that this effect was often discussed in the light of the "Psychology of the Uncanny" introduced by Ernst Jentsch in 1906. However, he, as well as others within the research community are moving away "from a psychological to a more existential-philosophical account of uncanny", which argues that the otherness within the self can provoke eeriness, especially when being confronted with humanlike technologies (Aydin, 2021, p. 305). The uncanny valley describes that human-like objects, as robots, might evoke emotional responses in humans, like real humans do (Misselhorn, 2009, p. 345). Positive and emotional responses are to be expected by human beings, the more human-like a robot or object is made. This is expected to be proportionate to the human-likeliness of the object. At a certain point comes a drop, where human will stop to perceive the object in a positive, empathic way and start to perceive eeriness instead. The positive annotations and empathy for the object, will only start to rise again, when the object comes indistinguishably close to a real human (Misselhorn, 2009, p. 346). Often this problem refers to aesthetic features, which ascribes a great role to embodiment again (ibid., p. 356; Schwind, 2015, p. 101).

The main characteristics for aesthetics, which also can be transferred to other expressions of human-likeness are typicality and salience. However, different features of the object may be irrelevant for empathy dependent on different kinds of emotions expressed or addressed. Especially, with AEI, where emotion recognition happens not exclusively visually, it is important to also consider features like voice, which can either sound very robotic and generated or genuinely human. Van Oosterzee (00:27:05) also mentioned that implementing emotion mimicking into AEI psychotherapeutic tools might have a great risk of resulting in the uncanny valley effect, which would not be desirable for therapeutical success. The underlying mechanism of how empathy with the object is created and how that turns into eeriness at a certain point is a phenomenon still under study.

Misselhorn's attempt (2009, p. 356) to explain the mechanism, from a philosophical perspective rather than an empirical, psychological perspective, describes that high human-likeliness triggers a certain concept so strongly, in this case empathy that people tend to fully apply it to the object. Since the object turns out to not be accepted as an instance of this concept, the attempt of application fails. Consequently, the process that evoked empathy is interrupted and no empathy is experienced anymore. This might lead to disenchantment of the object in the Weberian sense and people tend to be confused about why their attempt to apply the concept failed. This might provoke humans to perceive eeriness towards the object from this point on. As a consequence of the uncanny valley problem, AEI technologies must be designed in such a way that they are either not too human-like or simulate interaction with a real human in such a good way that it overcomes the uncanny valley and the related perception of eeriness, to ensure acceptance and positive, empathic annotations of the users of the AEI. Hanson (2006) was as optimistic as stating that an illusion of life, if possible, to create, might be able to mitigate the uncanny valley effect and "any level of realism can be socially engaging", when the aesthetics of a system are well designed.

To illustrate how these implications express themselves in and are approached or neglected in current attempts to develop AI-based tools for psychiatry, some examples will be assessed in the following section.

6.5 Examples of Current Developments and Applications

Although psychiatry is a very likely field of application of AEI, and a lot of research is conducted in this area, it seems as little is already implemented into real-world contexts and daily practice. This might explain, why most approaches to AEI in psychiatry are explorative by nature. So is the approach of de Mello and de Souza (2019, p. 1), who accepted the "challenge to develop an exploratory study on the nature of the combination of Psychology and Computation". Psychotherapy is a process that is exclusive to human beings, at least that is the traditional view. It is considered as being subjective and complex, however, parts of it can be computed (de Mello & de Souza, 2019, p.1). AI seems to be a useful tool when combined with human intelligence. Due to the lack of available psychological data, de Mello and de Souza (2019, p. 2) use a knowledge representation approach instead of a ML approach, so that psychology can benefit from computation in very specific tasks. The researchers stress right from the beginning that it is important to understand the limits of the approach they are using and see it as a method to assist in problem solving.

Psychotherapy is mainly concerned with raising awareness, self-awareness, and awareness about the condition of the patient, according to de Mello and de Souza (2019). Often people are reluctant to seek help, based on their economic and circumstantial situation, e.g., lack of time or money. Here AI systems might come into play as a resource of help. However, it cannot be neglected that they only are tools, which can be used as resources, aiming to support achievement of self-knowledge by the individual, provide useful information for therapeutic work and increase discernment. These tools are not curative in nature, they only serve to assist and improve effectiveness and quality of the psychotherapeutic work. This is in line with findings discussed before. The majority of the research society in this domain agrees that AI-tools should be used as assistive tools with human supervision and not work autonomously. They might make first contacts with and entering into therapy easier for potential patients. The proposed study aims for four objectives. First, it should investigate the contribution of AI in reconstruction work of possible (groups of) patients. Second, it should investigate how AI techniques and resources can be used to identify patterns of functionalities of possible patients. Third, the systemic link method, which is a method to explore the "patient's past to understand its history and identify unresolved issues" in three stages, should be used to identify historical bond and relational as well as affective patterns of potential patients (de Mello & de Souza (2019, p.4). Forth, the potential of AI in determining the overall patterns of couples and families by linking them to specific psychological patterns of individuals, should be explored. This would be what was predicted by Vogeley, de Bruin and partly Veldhuis as possibilities for AI in psychiatry, when they suggested to implement specific treatment approaches into the systems, which would be the systemic link method in this case.

The search for alternative solutions is reinforced by the establishment of extensive knowledge of the patient's history and personality, this can be automated by AI applications according to de Mello and de Souza (2019, p.3). Further, there were already attempts to use data mining to determine suicide risks and applications were developed that should serve as constant supportive companions during and after ambulatory treatment of people with clinical diagnosis. Further de Mello and de Souza (2019, p.4) argue for case-based reasoning in AI. Another process that can be automated by AI, is the anamnesis work. Usually, it takes around four sessions for the therapist to get to know the patient, this time can be reduced to offline sessions lead by AI via a pre-established set of questions. However, they warn to only use AI as an add-on resource for psychotherapeutic work. The evaluations and recommendations produced by the system need to be assessed by the therapist, which has also been shown is the common practice for AI-tools, since possible answers of chatbots are either reviews or pre-approved by therapists. AI-based anamnesis might reduce in-person sessions but does not replace them fully. Discussion in the sections before having already shown that some of the processes as anamnesis could be automated by AI. Practitioners as Veldhuis expressed interest in using AI to save some time by giving away some mainly bureaucratic and administrative work to have more time for actual patient interactions. Further, the great opportunities of data mining in this field were discussed before to have great potential to bring about new insights and breakthroughs in psychiatry. Nevertheless, there are still unresolved issues, as privacy concerns, which are neglected in the approach of de Mello and de Souza, as well has the high sensitivity of data used for psychiatric purposes. No sufficient solution has been found by now.

An example of an application designed for psychodiagnostical purposes is the virtual agent, Ellie. Ellie was developed by the project SimSei of the Institute for Creative Technologies in California. The platform involves several sensors and a webcam, which provide the input for facial recognition, movement, and voice analysis. Further, a complex system to process language and lead conversations is implemented. The interaction with humans is mediated by a virtual character, a woman in her thirties, who can be ethnically connected in various directions (Misselhorn, 2021, p. 74). This setup shows to potentially enable AEI and might potentially be able to establish empathic, compassionate relationships to the patients and to rule out ethnic bias, which were discussed in previous sections as problematic or hindering factors. The interaction happens in real time and questions and answers of Ellie are adapted based on the data gathered and processed by the system. Ellie is able to express verbally and nonverbally via mimic and gestures and her reactions seem empathic and benevolent. It is a multimodal system, which can express various emotions via the behaviour of the virtual agent. Users should feel comfortable and open up during conversation and Ellie should initiate situations, where verbal and nonverbal cues of psychological stress can be automatically analysed. If Ellie detects cues for psychological disorders, the system continues by asking deepening follow-up questions (Misselhorn, 2021, p. 78). A possible response for an evasive gaze of the human subject might be "I recognised that you were hesitant in answering, would you consider yourself to be a happy person in general?" At the beginning of the session, the system always clarifies that it is no psychotherapist, but wants to engage in conversation. This shows that the system is not meant to replace real human therapists and to not perceive the users. It states clearly that there is no human involved at this point in the conversation, which might make it easier for users to open up, as has been discussed before. People tend to be more open, when they do not fear being judged (Lovejoy et al., 2019, p. 2). However, in this case the illusion of a real human behind the technology, which was discussed as being effective before, is given away. However, this does not seem to lower the effect of the technology.

Trials on soldiers coming back from the war in Afghanistan showed that users are very willing to provide information. More symptoms of *Posttraumatic Stress Disorder* (*PTSD*) were disclosed by the soldiers than in conventional approaches aiming at diagnosing PTSD. This contradicts discussions mentioned earlier about not applying AI to those highly vulnerable patient groups. In the discussions before, it was mentioned to always use a human for those severe cases and only use AI with milder conditions. However, I think the positive outcomes speak for themselves and probably it is worth giving AI a chance for other cases. Again, it shows, that research in psychiatry and AI has been one sided so far and reduced to specific types of mental illnesses. Other types of mental illnesses might be investigated in research until a threshold for the use of AEI tools in psychiatry can be set. In general, users appraised Ellie very positively. A possible explanation, which was already sketched in the section about empathy is that Ellie was perceived as an empathic counterpart based on the implemented features. Additionally, the system was able to create a personal conversation atmosphere, which conveyed the impression of anonymity and being unobserved. The system presented itself as an unbiased counterpart that does not evaluate or judge the person (Misselhorn, 2021, p. 78). Hence, the technology is able to combine several factors that might reinforce successful treatment and tries to rule out and compensate as many hinderances as possible, e.g., biases, lack of empathy and privacy. Nevertheless, the technology was not knowingly applied outside testing settings, hence it cannot be said, whether issues as the uncanny valley effect might occur in real practice on the long run.

Other programs were designed to lead psychotherapeutic conversations, such as the ELIZSA program, which was able to imitate a written therapeutical conversation in such a convincing way that almost all patients felt understood, although they knew of the artificial nature of the system (Misselhorn, 2021, p.23). Abbou (2023, p. 108-111) introduces the chatbot Clare by clare&me, which is a voice-based AI that was developed to help with scarceness in psychotherapists. Users communicate via phone in form of a call with Clare, but also exchange via the WhatsApp messenger is possible. The current field of application is still subclinical and used for well-being purposes such as helping people with stress symptoms, which can use Clare to preventively reduce their symptoms. Research has shown that altercation and verbalisation of emotions already help to approach problems effectively. De Bruin also mentioned in the interview, that people already feel better, when they talk about their problems. The uncanny valley problem was also present here, although no aesthetics were involved. The VoiceBot first was created with a synthetic voice, but users refused that, when the voice was changed to a real human voice users started to accept the technology. To keep the distance and prevent users from becoming dependent on the VoiceBot, as it is often the case in therapeutical settings, it always clearly states that it is a VoiceBot, not a human counterpart. However, overreliance and difficulties to cut off consultation of the VoiceBot after some time could not be ruled out fully, so here further investigations and efforts to tackle the problem successfully are needed.

The company is planning to extent the abilities of the VoiceBot by e.g., measuring, and harnessing biosignals as voice or heartrate to improve diagnosis. This shows that AEI is slowly making its way into being implemented into psychiatric or at least well-being applications. Further, the company advertises with the slogan “No human, no judgement” and promising full anonymity, which, as has been shown in the case of Ellie, seem to be promising factors for success in the acceptance of users for psychotherapeutic tools. Abbou (2023, p. 109) adds that these factors might also lower the threshold of accepting such offers of help in a society, where mental and psychological health often is still considered as a subject of taboo. It was already discussed in previous sections, that AI may help to facilitate the entrance to therapy by ruling out stigma and taboo and its easy accessibility via already daily-used tools as smartphones.

An additional possible way of application of AEI for psychiatric purposes provides Microsoft, who bought the patent for a conversational chatbot that is modelled after specific real persons (Szewczyk & Janik, 2021, p. 38). This might offer an opportunity to possibly talk to dead persons and help with overcoming and managing grief. Vogeley addressed in the interview, that AI might be helpful to provide better solutions in the treatment of grief and loneliness, by e.g., avatar interaction. Also, an assisting function can be imagined for AEI applications in psychiatry. There is a burnout risk for highly sensitive doctors who are "very committed to helping their patients" (Szewczyk & Janik, 2021, p. 41). AI systems could help doctors to facilitate their emotions and thereby communicate better with the patients and maintain a healthy distance to professional matters. This likely might also apply for psychiatrists and psychotherapists, who likely might also experience suicide of patients at points in their career or oncological psychiatrists who accompany cancer patients and their families and friends along the way of the disease. Since future psychiatric and current well-being AEI tools, among other goals, aim to help patients and users with their emotion management and regulation, it is possible to equally assist practitioners with this task.

7. Discussion

After having shown some practical examples, I will now bring together, discuss, and set all results from the research conducted into a broader context. The research aimed at figuring out which abilities Artificial Emotional Intelligence could and should have when being applied in psychiatry, and which ethical and practical implications the implementation of AEI in psychiatry would have.

First of all, it has to be said that currently, any AI tools are rarely used in psychiatry, so AEI tools are even more rarely used. Hence, most of the research was fundamental and anticipatory work, that will help to set the stage for future research in this field. This research addresses a field which is currently understudied. While the body of literature on AI in psychiatry, is slowly growing, which was already noticeable during the half year, this research was conducted in, the use of AEI in psychiatry is still not sufficiently considered in literature. Hence, I will start off by discussing results for AI in psychiatry more generally, before moving to the specific implications of AEI, because if issues with AI in psychiatry are not solved, they will not allow moving further to applying AEI in psychiatry.

General implications of AI use in psychiatry were identified, which are governance, capitalisation, the mental health care system, data handling, the application in a consequential setting, domains of application and responsibility. Based on the expert interviews and findings from the scientific community, it can be said, that governance of AI applications in psychiatry is required. Despite, there was no consensus or clear line of governance provided, it can be concluded, that successful governance should protect the rights of patients and practitioners while enabling to use the potential AI is offering for improving psychiatric practice. This includes preventing capitalisation, which was often raised as a concern by the interviewees and the scientific community. I agree that it is not ethical to make profit out of the vulnerabilities of persons, especially, when they are dependent on the products offered to improve their mental and psychological condition. The quality and safety of the products should always be valued over personal profit. Solutions offered to ensure this are e.g., convincing insurance companies to support products and cover the costs for the patients within their service (Vogeley), introducing governmental labels to hint which mental health applications are scientifically validated and safe to use (Lovejoy et al., 2019, p. 2) and restricting or banning AI technologies until (ethical) concerns and negative implications are ruled out for the use in psychiatry (Katirai, 2023, p. 8).

I think these are valid suggestions for the possible governance of AI in psychiatry. However, I think a full ban on AI in psychiatry might hinder progress and lead to stagnation in research. There is already little to no practical experience available on AI in psychiatry and a ban might lead to researchers refusing to further invest time and resources in exploring the potential of AI in psychiatry. The examples provided have shown that it is worthy to pursue research in the field of psychiatry since AI applications as Ellie have already contributed to great outcomes, that were outperforming traditional methods, e.g., in diagnosing PTSD ((Misselhorn, 2021). It is important to mention that this application was developed in the US, which has different regulations than e.g., the EU and might therefore be freer in research while neglecting other important implications. One of them might be privacy.

Privacy is an ethical concern that was also addressed by some of the experts (de Bruin; van Oosterzee) to be blocking research and not being as important as other values. However, I tend to agree with other experts and the overall scientific community that privacy should not be neglected and has its validation within the debate. Sure, a weighing of values should always be considered and fundamental values such as preventing death should always be the highest priority, but fully neglecting other values to achieve this cannot be the solution. Other ways need to be found to facilitate this debate in order to satisfy all parties. A trade-off or compromise might be required, but how this will look like is not clear yet. It is advisable to find a global or at least European solution for the governance and regulation of AI in psychiatry, as it is the aim of the AI Act. This will ensure unity in regulation across a broader region since AI applications are likely to spread across national borders. Governing privacy includes finding an ethical way of data handling. Processing data anonymously, as handled within the chatbots Ellie and Clare, might help to protect privacy, and still ensure data to be stored and used for research purposes that might contribute to creating a new knowledgebase for the field of psychiatry. This was discussed as being a huge advantage of the new possibilities enabled via AI in psychiatry by experts as de Bruin and van Oosterzee.

Another important point was that AI in psychiatry means, that AI is implicated in a consequential setting, it has immediate implications on the users, in this case often vulnerable patients. The interaction with the psychiatric AI tool leads to responses of the users and changes within the users. Often these changes are the purpose of the interaction, as when AI is meant to assist in therapy and with emotion management. Here behavioural change or change to another emotional state are provoked by the intervention of the AI system. This is required for therapy, however, it is not unproblematic, because it can also be interpreted as manipulation, which would be considered as ethically critical (Schuller & Schuller, 2018, p. 40). This is especially the case when it comes to AEI, because the systems do not have emotions or are emotional themselves, as has been figured out. All they can do is mimic emotions for more authentic and empathic interactions. However, these interactions will never be real, but ever be an illusion or perception of emotions. The systems simply use the data fed into the systems by humans and either generate a response that is very human-like or use a pre-programmed response created by humans as a response, that creates the impression of emotional interaction and commitment. However, these emotional sounding responses are not a reflection of the systems inner perception or feelings, because it has none. The responses simply trick the counterpart into thinking that it has some and this is ethically not correct.

Although a lot of systems, as shown in the examples (Abbou, 2023; Misselhorn, 2021) clearly state that they cannot feel what the human is feeling and telling them, since they are no human counterparts, but machines, patients feel understood and empathised with despite this is beyond the capacity of the system and merely an illusion created. One would think that humans are aware that the output of the AI systems are not real, but simply generated responses, but that is enough to evoke real responses and emotions within them. Consequently, the artificial emotional interactions cause real responses in implications for the human user, who is as a patient in a psychiatric setting not necessarily in the state to reflect sufficiently on the artificiality of the interactions, since the patient is in a vulnerable state. I think human oversight is required and the responsibility should be clearly assigned to a human actor, be it a supervising human therapist, the company or a neutral third party. Artificial emotional interactions might have promising good implications and healing impacts on the patients; however, this is not guaranteed as has been discussed before. Using technology is no success guarantee, everyone reacts differently to the technology and as some unfortunate examples of e.g., suicide after ChatGPT's interventions have shown (Hale, 2023), there needs to be human supervision to prevent such cases from happening and have a responsible party involved, who can step in.

Due to these issues the domain of applying AI in psychiatry is also a fair point in this debate and it has been shown that decisions about this are not that easy. I agree with the experts Vogeley and Veldhuis, that an application for patients, suffering from milder mental health conditions, who still have resilience, sounds most promising. The implications for people with more severe conditions as just discussed might be harmful if they cannot deal sufficiently with their conditions by mere AI-intervention. However, I still think it is very difficult to set a fixed threshold on when to use AI-tools and when not to use them. I think it cannot be set for specific conditions and a decision should rather be made on a case-to-case basis. If you exclude certain mental health conditions per se, it might prevent AI to unfold its full potential, as has been shown in the example with PTSD. It was suggested to not use AI tools with PTSD patients, because of their severe conditions and to show real human respect. At the beginning of my research, I would have fully agreed with this claim, because I thought those people are too vulnerable and too sensitive topics are involved to be treated simply by technology. However, experiments have shown that AI tools were even more successful in collecting information from veterans suffering from PTSD than humans ever were. Hence, I now think the decision on when to use and when not to use AI in psychiatry should not be generalized and every patient should get the opportunity to be helped by AI tools, when they wish to and when the tools reach the state of being valid and safe to use.

Despite all the possibilities AI is promised to offer, the mental health care system still needs to change. This was suggested by almost all experts. I agree that AI should not be used as a techno-fix to solve problems quickly. Sure, AI might be helpful to build a bridge into therapy for people, who are afraid of entering first, because it is taboo in their social environment to go to therapy, or they are afraid of first human contact with a psychiatrist or therapist and fear to be judged or because there is simply no human professional available. AI might serve as a first neutral contact to reach out to and sometimes having the impression someone is listening might already be enough to improve the situation for a person. Nevertheless, at a certain point a human professional should enter the process and facilitate further interactions and treatments. Using AI should not be an excuse to not tackle general issues of the mental health care sector, as e.g., the scarcity of professional practitioners, the difficulties to enter the mental health care system and the high costs some patients simply are not able to cover.

Additionally, on the other side, the health care system needs to change and open up in order to allow and enable AI to enter the field of psychiatry in practice. As de Bruin mentioned, the conservative community of psychiatry needs to change their values and be more open to allow being convinced by the possibilities and positive, promising implications AI might add to the field. Requirements, regulations, and conditions of the mental health care sector need to be adapted to create the preconditions for AI entering the practice. Moreover, the AI tools need to reach maturity, validity, and safety to be able to be applied in the field without causing harm and other possible negative implications.

Implications that explicitly involve emotions and AEI include realism, over-reliance on technology, measurements and definitions of emotions and mental conditions, transferability to real-world, artificial empathy, compassionate interactions, the uncanny valley effect, and the black box problem. Right at the beginning, when I set the theoretical foundation for my research, it figured out, that the whole field of emotions and consequently also Artificial Emotional Intelligence has a substantial problem, because it is lacking a universal conception of emotions. It was discussed that the field is lacking consensus on how to define emotions and how to measure them. Many attempts across diverse disciplines, including philosophy and psychology, have been made to define the concept of emotions and capture every aspect of it. However, no universal, shared definition was found. Consequently, it was decided to present the definition, which is commonly used as the basis for research and development in the field of affective computing and AEI. This is the common view, or the basic emotion approach, based on the work of Ekmann.

I could not neglect, that this approach has undergone harsh criticism to be too limited and not able to capture what emotions really are and I agree that it is not good practice to simply reduce the complex concept of emotions to six basic emotions and ignore a history of research that has happened after the establishment of this approach and still reply on this reductionistic view. Valid points of criticism, which were proven to be scientifically valid among others by Barrett and colleagues (2023) are, that emotions are highly context dependent and subjective. As the three main shortcomings limited reliability, lack of specificity and limited generalizability were identified. The conceptualisations of emotion are often not inclusive and neglect variety in emotion perception, expression, recognition, and interpretation among cultures and even individuals. Hence, real-world applications are often not possible beyond laboratory settings. Further, clear markers and measurements are lacking to identify emotions. The results and definitions often do only apply to the WEIRD nations (Katirai, 2023).

This is also problematic, because the lack of a universal definition of emotions implies that all AEI technologies, based on the chosen outdated definition of the common view, lack substantial scientific foundation. This makes their application either limited to WEIRD nation's contexts, or other contexts they were developed in, but they cannot be applied in any context without limitations in reliability and validity. Probably, their results are fully invalid for the context they are applied in because the understanding and expression of emotions is way different than implemented into the technology. Hence, the first and most important issue to tackle for AEI is to find a universal definition for emotions. Since this might not be possible at all, as the years of debate has shown, the definition underlying such technologies should at least be more inclusive and adaptive to variations, users, and contexts. Literature and experts addressed the lack of consensus for definitions and measurements also for other related topics as empathy, well-being, and even mental illnesses, as the currently used classification systems seem to be outdated as well. Here, the same applies, a sufficient common ground needs to be found to pursue research and development of related technologies, especially AEI, responsibly and create safe and valid products.

Here the black boxed nature of AI can be transferred to emotions, both processes are not explainable yet and hence often cannot be fully validated. Since the data used to train AEI for psychiatric purposes is often invalid and untested, the situation is even worse in this case. Transparency cannot be guaranteed, and it is likely that the AEI will pursue biases and stereotypes. I think this is highly unethical since psychiatry is a field that already has to fight a lot of negative implications stemming from the pursuit of stereotypes, especially regarding specific diagnoses. Since the research on AI in psychiatry currently is rather one-sided and only few mental illnesses are studied frequently, the database of the AI systems cannot be as diverse as needed. A lot of mental conditions are still underrepresented as well as all the emotion data that is difficult to be used outside the Western industrialized world. This implies that research is needed to make the systems as diverse and integrative as possible before they move to practice. Nevertheless, experts and members of the scientific community (Abrams 2023; de Bruin) think that these issues are solved at some point and AI’s performance will be convincing enough to not require full understanding and transparency anymore. I cannot fully agree with this statement because I think issues as bias should be resolved responsibly and therefore vanish from the discussion and not only because they got outweighed by potential benefits.

Another factor is, that the most common and advanced application for AEIs is that they are only able to recognise emotions so far. In the best cases, AEI tools are able to respond to human emotions, this is most likely for conversational agents (Schuller & Schuller, 2023, p. 41). Consequently, especially for psychiatry the most likely application will be chatbots, who are also the most common application under development. There is a whole repertoire of chatbots already available, however, still only in well-being contexts since they are not mature enough to be safely applied in psychiatric context. Nevertheless, chatbots might be the application of the future in psychiatry, and almost all experts, as well as researchers of the scientific community mentioned them as possible application.

I think it will be a great advantage, that these chatbots seem to be neutral, non-opinionated and non-judgemental conversation partners, as was argued in the literature and shown by some examples (Lovejoy et al., 2019, p. 2; Misselhorn, 2021, p. 78). As already discussed, this will help a lot to make people more confident in engaging in therapy. However, some other issues need to be resolved to improve the users experience and make AEI a safe and valid tool to use in psychiatry. Despite expertise in the field of psychiatry the atmosphere seems to be very important in a therapeutic setting. A factor that was especially stressed by many experts in the interviews and scratched in the literature (Misselhorn, 2021) is the relationship between psychotherapist and patient. Although, van Oosterzee mentioned, that empathy and emotions are not necessary to be included in therapy from the therapist’s side, I disagree and think that it is still necessary to be empathic and show understanding as a therapist, be it an artificial agent or a human one, to validate patients’ emotions during therapy. To establish this relationship, it was argued that empathy is required, and technology is often lacking this. Misselhorn (2021) introduced the concept of artificial empathy, where attempts were made to create empathy by either implementing theories or emotion data into the AI systems. Van Oosterzee raised concerns that this method might lead to information loss and oversimplification, which I agree with. Since it is likely that other approaches to emotions besides empathy are implemented into AEI with similar strategies, there should be experts involved, who are rooted in the field the theory is derived from to ensure the loss to be as minimal as possible and to validate that the model still represents what it was intended to.

Bollen argued that these attempts at least create the impression of an empathic response or interaction for the user. However, it will always be a representation of the level of empathy integrated by the developers and never be an empathic technology at all. As discussed before, I think this will be a form of manipulating the users and vulnerable patients, which is ethically questionable to pursue. Nevertheless, examples as Ellie and Clare have shown, that users feel empathized with and feel understood. Consequently, some attempts to implement empathy into chatbots have shown to be successful and the systems convey emotions as intended. They were able to create a trusting, comfortable atmosphere that made the patients disclose intimate, sensitive information needed for diagnosis and treatment (Misselhorn, 2021). This likely would not have been the case if the patients would not have perceived a warm connection and compassion being conveyed from their conversation counterpart. As Veldhuis described, if this condition is given and AEI is able to establish or at least create the impression of such relationships, the willingness to engage and seek help from the system will increase within the patients. Hence, I think, if some more effort is made to create an authentic impression of empathy, it can be a promising tool with helpful implications for patients and therapists in psychiatry. I do not think that patients will bother whether it is real human empathy unless they feel understood and receive help. As de Bruin argued, it might be enough for them to get the impression there is a professional behind the tool who takes care. Regardless of this, I still think human supervision is required and a human agent should be involved, who is responsible and checks regularly for negative implications caused within the patient by the system.

If these warm relationships can be established via AEI interaction, overreliance will get a serious issue as discussed. Patients will likely get too attached and dependent on their artificial therapist and rely too heavily on its all-day-long availability and advice. This issue was addressed by the experts and in the literature frequently. Hence, I think it is a serious issue. Research about gaming addiction can possibly be used to transfer knowledge to this new phenomenon in order to better approach it. Since it is already difficult for human therapist to find the right point of making a cut and ending therapy, this decision cannot be left with the AI. Consequently, I agree with the experts and voices in the literature, that AEI should only be used as an add-on in therapy. It should not be applied autonomously and only serve as an assisting tool for practitioners helping out or providing additional support during times the human therapist is unavailable.

The argument that effective expressions of machines might not be convincing enough, provided by Pashevich (2021), which leads to loss of empathy was only raised in the context of robots so far. Since AEI not necessarily is embodied, one might think that it does not apply to AEI. However, similarly as the discussion of the uncanny valley effect has shown, which usually also is linked to aesthetics, the effect can also occur in relation to other human features as voice. In the case of Clare, a synthetic voice was disliked by the users, which caused a change to a human, recorded voice. I think this can be explained by Pashevich's line of argumentation as well as by the uncanny valley effect itself. Especially, for the case of including emotions the effect will occur around other factors than aesthetics, it is likely that will also occur, e.g., when a conversation feels too naturally human, but as it is the practice currently, the AEI will disclose and remind the user at certain points that it is a machine and not a human counterpart. These situations might disenchant the AEI during the conversation and case eeriness within the users. Although this has not been shown yet, it might occur when AEI is more integrated into the daily practice of psychiatry.

7.1 Limitations of the Research

This research does not come without some limitations. First, it has to be mentioned that the AEI in psychiatry is a relatively new topic and not much literature is out there to study. AI in general is only rarely used in psychiatric practice, which made it almost impossible to find some actual real-world examples. Because of this the initial intention of the expert-interviews to gain practical insights of practitioners working with AEI in psychiatry needed to be changed and a more speculative and anticipatory approach was chosen, where the experts could transfer knowledge from practice in related fields, e.g., e-health technologies, to estimate potential implications of AEI in psychiatry. Additionally, experts from the research domain were included who could provide insights into which direction the developments are progressing within this field. It also has to be noted that the network of experts is rather small and only includes persons from a similar research domain, who share similar perspectives and values. Including experts from different parts of the world and more diverse backgrounds might have led to results representing a broader population.

The perspective of developers and companies is missing. Attempts were made to approach possible interview partners who are involved in the development or work for a company that is providing e-health products for psychiatry, to also have insights from the provider's perspective and possibly gain insights into the internal workings of such technologies. However, no one responded. In general, the response rate of possible interview partners was rather low, which prevented the gathering of more diverse opinions from a broader sample. Nevertheless, the interviews provided a great source for insights into potential practice and practical research and were helpful in identifying potential implications of AEI technologies in psychiatry. They also disclosed information that might be perceived as uncomfortable truths, which are often held back in scientific publications.

Another limitation of this research was that all results are dependent on the definition of emotions chosen at the beginning. This might lead to a narrower focus and that some issues were not addressed. Additionally, results could be interpreted differently if another definition was chosen. Since the consensus about a definition is very divided across the scientific community, the definitions are not used unified and every discipline has its own approach, choosing a definition was necessary to build a frame and point of reference for this research. Overall, this research contributed to build a basis for future research, because it started to bring together information that will help to explore the field of AEI in psychiatry and will slowly help the fill the gap, which currently exists in literature.

Another limitation is the limited availability of information and data about specific AEI applications and how they are developed. This prevented to explore specific programs into depth and gather insights into the data used to develop the systems. Consequently, it was not clear which conception of emotion was used for the systems presented in the examples and it needed to be assumed that the common view of emotions was applied. However, this issue at the same time is a result, which stresses the importance of transparency for AEI in psychiatry. Systems currently cannot be labelled as valid, reliable, and safe, unless the underlying theoretical basis and data are examined. Since there was no information available, companies need to improve transparency and enable third parties to check their systems for safety and validity.

Further, there are many more implications of applying AEI in psychiatry, especially ethical ones, which could not be captured in this research, because it would have exceeded the scope of this work to include all of them. Since all practical implications cannot be fully identified yet, it is advisable to constantly keep track and map them when occurring during development and testing phases to be aware of them before implementation. There also is room for investigating in further ethical implications, as e.g., the issue of anthropomorphisation and mystification of AEI technologies.

8. Conclusion

To sum up, several steps were taken to answer the research question “Can and should an Artificial Intelligence be emotionally intelligent especially within the field of psychiatry and what ethical and practical implications might an emotionally intelligent Artificial Intelligence have in psychiatry?”. First the current status of research was assessed, identifying gaps in the scientific literature to fill. The main aspects were the lack of definitions and measures for emotions and the lack of concrete application of AEI technologies in psychiatry because of unresolved ethical issues. After that important theoretical concepts were discussed to provide a common ground. Here the first results addressing the research question and thesis could be drawn. The lack of a universal definition of emotions poses challenges to reliably implement emotional intelligence into AI systems. Hence, all the results of my research need to be interpreted in the light of the presented definition of emotions. Further, it can be said, that a part of the research question can be answered by saying that whether AI can be emotionally intelligent is dependent on the definition and context used.

Next, expert interviews were conducted to ensure the practical applicability of the results of this research and be able to figure out possible practical implications of AEI in psychiatry. After that further practical and ethical implications were identified and discussed using the scientific literature. Referring to the research question it can be said that practical implications of AEI in psychiatry among others include: improving the general mental health care system to enable using AI’s promised advantages without neglecting to tackle the issues at its root, the domain of application, where positive implications especially are expected for patients with milder conditions, and privacy and data handling, which at the same time is an ethical implication. Privacy and data handling can be partly tackled by asking for consent and anonymising the data to still be able to use the new sources of data and knowledge production opened up by AEI. Nevertheless, a sufficient solution needs to be found for this implication and sufficient governance might be a possible solution.

Governance is also needed to tackle further ethical implications as unethical capitalisation of AEI technologies, especially in vulnerable psychiatric settings, protecting and ensuring patients’ rights and safety by e.g., labelling AEI mental health applications as reliable and safe to use. Further, overreliance on AEI tools in psychiatry can be tackled by making human oversight a requirement and only allowing them as a supplement to human therapy. The lack of a common scientific basis for the definition and measurement of emotion still leads to invalid tools. Hence, this issue needs to be tackled first and the development of a more adaptive and inclusive account of emotions is required that is not limited to WEIRD nations' perspectives. Implications of adding emotionally intelligent skills to AI technologies, as chatbots, might have positive implications, especially for trusting relationships between patient and AEI agent which might reinforce successful therapeutic interactions. Nevertheless, these interactions might be overshadowed by the uncanny valley effect. If this effect can be facilitated sufficiently, AEI interaction might be perceived by patients as engaging and compassionate. However, patients need to be sufficiently education about the AEI tools and be aware, that these emotionally and empathically perceived interactions still are only mimicry and always of artificial nature, so they should be aware of possible manipulation. To successfully implement AEI, governmental regulations are necessary until negative ethical and practical implications are ruled out and the technologies are safe to use.

Finally, to answer if an Artificial Intelligence can and should be emotionally intelligent in psychiatry, it can be said that AEI in psychiatry can have the emotionally intelligent skills of emotion recognition and mimicking empathy and compassion during social interactions with patients. It can be concluded that AEI should have these skills to make the application in psychiatry more ethically acceptable, while still being cautious that the implementation of these skills adheres to certain safety and ethical standards to prevent harmful implications. Hence the thesis can be partly confirmed, since AI can by now only have some skills of emotional intelligence and is limited in learning others. I think the emotional intelligence skills discussed, would make the application of AI in psychiatry more acceptable when it can create a trusting, less judgemental atmosphere for the patient. Additionally, several positive and negative ethical and practical implications come along with adding emotional intelligence to AI, which are summarised above. Further, the specific implications of (emotional) AI in psychiatry highly depend on the definition of emotions and the AI application or tool used.

As an outlook for future research, I might suggest investigating in identifying further ethical implications and find solutions for the existing ones. Moreover, a fundamental, scientific, adaptive, and inclusive common ground for defining and measuring emotions should be found. This is crucial to be able to create valid and safe applications for the mental health sector. Research should also be conducted more diversely within the field of AI and psychiatry, to make applications more applicable to a wider range of mental illnesses and not be limited to a few specific ones anymore. This could help to unravel the full potential safe and responsibly designed AEI tools might have in psychiatry.

9. Literature

Abrams, Z. (2023, July 1). AI is changing every aspect of psychology. Here’s what to watch for. *Monitor on Psychology*, *54*(5), 46.

Allen, S. (2022, November 3). *Improving psychotherapy with ai: From the couch to the keyboard*. IEEE Pulse. https://www.embs.org/pulse/articles/improving-psychotherapy-with-ai-from-the-couch-to-the-keyboard/.

André, E. (2023, March 24-26). *Roboter als empatische Gegenüber des Menschen? Die Bedeutung emotionaler KI für die Mensch-Maschine-Interaktion*. [Conference presentation]. Roboter als empatische Gegenüber? Emotionale KI und menschliche Freiheit, Loccum, Germany. https://www.loccum.de/tagungen/2313/.

Abbou, K. A. S. (2023). *Menschenversteher:* Wie emotionale Künstliche Intelligenz unseren Alltag Erobert. Droemer Verlag.

Ackley, D. (2016). Emotional intelligence: A practical review of models, measures, and applications. Consulting Psychology Journal: Practice and Research, 68(4), 269-286. https://doi.org/10.1037/cpb0000070.

Aday, J., Rizer, W., & Carlson, J. M. (2017). Neural mechanisms of emotions and affect. *Emotions and Affect in Human Factors and Human-Computer Interaction*, 27–87. https://doi.org/10.1016/b978-0-12-801851-4.00002-1.

Albraikan, A. A., Alzahrani, J. S., Alshahrani, R., Yafoz, A., Alsini, R., Hilal, A. M., Alkhayyat, A., & Gupta, D. (2022). Intelligent facial expression recognition and classification using optimal deep transfer learning model. *Image and Vision Computing*, 104583. https://doi.org/10.1016/j.imavis.2022.104583.

Asada, M. (2015). Towards artificial empathy. *International Journal of Social Robotics*, *7*(1), 19–33. https://doi.org/10.1007/s12369-014-0253-z.

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, *22*(6). https://doi.org/10.2196/15154.

Aydin, C. (2021). The technological uncanny as a permanent dimension of Selfhood. *The Oxford Handbook of Philosophy of Technology*, 298–317. https://doi.org/10.1093/oxfordhb/9780190851187.013.18.

Baltrusaitis, T., McDuff, D., Banda, N., Mahmoud, M., Kaliouby, R. el, Robinson, P., & Picard, R. (2011). Real-time inference of mental states from facial expressions and upper body gestures. *Face and Gesture 2011*. https://doi.org/10.1109/fg.2011.5771372.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1),168. https://doi.org/10.1177/1529100619832930.

Bar-On, R. (1997). Technical manual for the Emotional Quotient Inventory. Toronto, Ontario, Canada: Multi-Health Systems.

Bergstrom, C. T., & West, J. D. (2020). *Calling Bullshit: The Art of Skepticism in a Data Driven World*. Random House.

Birks, Y. F., & Watt, I. S. (2007). Emotional intelligence and patient-centred care. *Journal of the Royal Society of Medicine*, *100*(8), 368374. https://doi.org/10.1258/jrsm.100.8.368.

Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., Loreggia, A., Murgesan, K., Mattei, N., Rossi, F., & Srivastava, B. (2021). Thinking Fast and Slow in AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 15042–15046. https://doi.org/10.1609/aaai.v35i17.17765.

Bostrom, N., & Sandberg, A. (2011). *The Future of Identity*. Report, Commissioned by the UK's Government Office for Science.

Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry*, *218*(3), 131–134. https://doi.org/10.1192/bjp.2019.245.

Butnariu, M., & Sarac, I. (2019). Biochemistry of Hormones that Influences Feelings. *Pharmacoepidemiology and Drug Safety*, *1*(1), 1–6.

Caruso, D. (2003, November). Defining the inkblot called emotional intelligence. Retrieved from http://www.eiconsortium.org/reprints/ei_issues_and_common_misunderstandings_causo_comment.html.

Cohen, S. (2021). The Basics of Machine Learning: Strategies and Techniques. *Artificial Intelligence and Deep Learning in Pathology*, 13–40. https://doi.org/10.1016/b978-0-323-67538-3.00002-6.

Cooper, R. (2004). What is Wrong with the DSM? *History of Psychiatry*, *15*(1), 5–25. https://doi.org/10.1177/0957154X04039343.

Czerwinski, M., Hernandez, J., & McDuff, D. (2021). Building an AI That Feels: AI systems with emotional intelligence could learn faster and be more helpful. *IEEE Spectrum*, *58*(5), 32–38. https://doi.org/10.1109/mspec.2021.9423818.

de Mello, F. L., & de Souza, S. A. (2019). Psychotherapy and artificial intelligence: A proposal for alignment. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00263.

Daily, S. B., James, M. T., Cherry, D., J. Porter, J., Darnell, S. S., Isaac, J., & Roy, T. (2017). Affective computing: Historical foundations, current applications, and future trends. *Emotions and Affect in Human Factors and Human-Computer Interaction*, 213–231. https://doi.org/10.1016/b978-0-12-801851-4.00009-4.

Döringer, S. (2020). 'The problem-centred expert interview'. Combining qualitative interviewing approaches for investigating implicit expert knowledge. *International Journal of Social Research Methodology*, *24*(3), 265–278. https://doi.org/10.1080/13645579.2020.1766777.

Frajo-Apor, B., Pardeller, S., Kemmler, G., & Hofer, A. (2015). Emotional intelligence and resilience in mental health professionals caring for patients with serious mental illness. *Psychology, Health &amp; Medicine*, *21*(6), 755–761. https://doi.org/10.1080/13548506.2015.1120325.

Gebhard, P. & Inthorn, J. (2023, March 24-26). *Emotionale KI: Was kann und was soll sie leisten? Und welche Fragen stellen sich aus theologischer Perspektive?*. [Conference session]. Roboter als empatische Gegenüber? Emotionale KI und menschliche Freiheit, Loccum, Germany. https://www.loccum.de/tagungen/2313/.

Goleman, D. (1995). Emotional intelligence: Why it can matter more than IQ. New York, NY: Bantam Books.

Goleman, D. (1998). Working with emotional intelligence. New York, NY: Bantam Books.

Grabowski, K., Rynkiewicz, A., Lassalle, A., Baron-Cohen, S., Schuller, B., Cummins, N., Baird, A., Podgórska-Bednarz, J., Pieniążek, A., & Łucka, I. (2018). Emotional expression in psychiatric conditions: New technology for clinicians. *Psychiatry and Clinical Neurosciences*, *73*(2), 50–62. https://doi.org/10.1111/pcn.12799.

Hale, E. (2023, April 27). *CHATGPT is giving therapy. A mental health revolution may be next*. Technology | Al Jazeera. https://www.aljazeera.com/economy/2023/4/27/could-your-next-therapist-be-ai-tech-raises-hopes-concerns.

Hanson, D. (2006). Expanding the aesthetic possibilities for humanlike robots. In Proc. IEEE Humanoid Robotics Conference, special session on the Uncanny Valley, Tskuba, Japan.

Herrando, C., & Constantinides, E. (2021). Emotional contagion: A brief overview and future directions. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.712606.

Hofmann, B. (2015). Suffering: Harm to bodies, Minds, and persons. *Handbook of the Philosophy of Medicine*, 1–17. https://doi.org/10.1007/978-94-017-8706-2_63-1.

Katirai, A. (2023). Ethical considerations in Emotion Recognition Technologies: A review of the literature. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00307-3.

Kumar, H., & Martin, A. (2022). Artificial Emotional Intelligence: Conventional and Deep learning Approach. *Expert Systems with Applications*, 118651. https://doi.org/10.1016/j.eswa.2022.118651.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2.

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*(10). https://doi.org/10.1073/pnas.1619316114.

Lovejoy, C. A., et al. (2019). Technology and mental health: The role of Artificial Intelligence. *European Psychiatry*, *55*, 1–3. https://doi.org/10.1016/j.eurpsy.2018.08.004.

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Selecting a measure of emotional intelligence: The case for ability scales. In R. Bar-On & J. D. Parker (Eds.), Handbook of emotional intelligence (pp. 320 –342). San Francisco, CA: Jossey-Bass.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), Emotional development and emotional intelligence: Implications for educators (pp. 3–31). New York, NY: Basic Books.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). Toronto, Ontario, Canada: Multi-Health Systems.

McDuff, D., El Kaliouby, R., Kassam, K., & Picard, R. (2010). Affect valence inference from facial action unit spectrograms. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. https://doi.org/10.1109/cvprw.2010.5543833.

Michalska, K. J., & Davis, E. L. (2018). The psychobiology of emotional development: The case for examining sociocultural processes. *Developmental Psychobiology*, *61*(3), 416–429. https://doi.org/10.1002/dev.21795.

Misselhorn, C. (2009). Empathy with inanimate objects and the Uncanny Valley. *Minds and Machines, 19*(3), 345–359. https://doi.org/10.1007/s11023-009-9158-2.

Misselhorn, C. (2021). *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co.. was Bedeutet das Alles?* Reclam, Philipp.

Ortony, A. (2022). Are all “basic emotions” emotions? A problem for the (basic) emotions construct. *Perspectives on Psychological Science*, *17*(1), 41–61. https://doi.org/10.1177/1745691620985415.

Pashevich, E. (2021). Can communication with social robots influence how children develop empathy? best-evidence synthesis. *AI & SOCIETY, 37*(2), 579–589. https://doi.org/10.1007/s00146-021-01214-z.

Powell, K. R., Mabry, J. L., & Mixer, S. J. (2015). Emotional intelligence: A critical evaluation of the literature with implications for Mental Health Nursing Leadership. *Issues in Mental Health Nursing*, *36*(5), 346–356. https://doi.org/10.3109/01612840.2014.994079.

Salovey, P., Mayer, J. D., & Caruso, D. (2002). Emotionally Intelligent Certification Workshop. Toronto, Ontario, Canada.

Schwind, V. (2015). Historical, cultural, and aesthetic aspects of the Uncanny Valley. *Collective Agency and Cooperation in Natural and Artificial Systems,* 81–107. https://doi.org/10.1007/978-3-319-15515-9_5.

Shargel, D. (2016). Appraisals, emotions, and inherited intentional objects. *Emotion Review*, *9*(1), 46–54. https://doi.org/10.1177/1754073916658249.

Sullivan, E. (2020). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 000. https://doi.org/10.1093/bjps/axz035.

Schuller, D., & Schuller, B. W. (2018). The Age of Artificial Emotional Intelligence. *Computer*, *51*(9), 38–46. https://doi.org/10.1109/mc.2018.3620963.

Szewczyk, R., & Janik, K. (2021). How Much Emotionally Intelligent AI Can Be? In *Control, Computer Engineering and Neuroscience* (p. 37–49). Springer International Publishing. https://doi.org/10.1007/978-3-030-72254-8_5.

Tracy, J. L., & Randles, D. (2011). Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review, 3*(4), 397 405. https://doi.org/10.1177/1754073911410747.

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0.

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on Affective computing: Emotion models, databases, and recent advances. *Information Fusion*, *83–84*, 19–52. https://doi.org/10.1016/j.inffus.2022.03.009.

Wilson, C. (2014). Semi-structured interviews. *Interview Techniques for UX Practitioners*, 23–41. https://doi.org/10.1016/b978-0-12-410393-1.00002-8.

Wolff, S. (2005). Emotional Competence Inventory: Technical manual. Boston, MA: The Hay Group.

World Health Organization. (2023, February 6). *Artificial Intelligence in mental health research: New who study on applications and challenges*. World Health Organization. https://www.who.int/europe/news/item/06-02-2023-artificial-intelligence-in-mental-health-research--new-who-study-on-applications-and-challenges.

Appendix A

Interview Guide English

Explanation of goals and purposes of research

Artificial Intelligence (AI) became a very prominent topic currently. Its use and application entered almost every field from business to leisure activities. It seems that one can no longer avoid contact with AI during daily life. Research rapidly continues and AI’s possibilities seem to be endless and beyond imagination. AI is often built upon human principles and with the aim to become at least as good at humans in certain tasks, if not outperform them. Hence, it is not surprising that research is also happening in the direction of emotions. Artificial Emotional Intelligence (AEI) is one of the visions for the future. The aim is to develop an AI that can recognise, interpret, process, and simulate emotions. This enables more individual approaches and interactions. The aim of my work is to find out which ethical and practical implications an emotionally intelligent AI might have. I am exploring the implications, especially with a focus on the use-case of Artificial (Emotional) Intelligence within psychiatry/psychotherapy.

I want to explore the benefits and challenges AEI adds and poses to those disciplines and how AEI will change the practice and understanding of those discipline of psychiatry/psychotherapy. By carrying out expert interviews in the field, I hope to gather insights into specific (possible) practical implications of the use of AEI in psychiatry and psychotherapy.

Asking for consent on recording the conversation. The conversations will be stored securely.

The interview will take approximately 45 minutes.

Themes	Interview Questions
General information and work setting	Can you introduce yourself and what links you to EAI and psychiatry/ psychotherapy?
General EAI (applications)	What do you know about the **current stage of AEI and how it is used within psychiatric/psychotherapeutic practice**? Possible Follow-up Questions:

	Which **specific applications** do you know? Please tell me more about application xy. Do you use AEI applications/technologies in your daily practice? - If no, would you like to? If no, why not? - If yes, which ones and why? Are you involved in the development/research or testing of AEI technologies/applications? If yes, please tell me more about it.
Possibilities and Challenges	Now that we talked about some existing AEI technologies/applications, let's talk a bit about where you see the **most potential** that AEI technologies might add to the psychiatric/psychotherapeutic practice. What do you think, in which areas or for which purposes will AEI be **beneficial and why?** Of course, there are not only positive views and visions surrounding this topic. Where do you see the **challenges** coming along with introducing AEI into Psychiatric/psychotherapeutic practice? What other challenges (besides ethical/technical ones) do you see/expect? What are **your visions for your/the future** research on AEI in psychiatric/ psychotherapeutic practice? Where do you

	think you might face difficulties/challenges and why?
Desirable or not?	Do you think AEI is a desirable tool to implement into psychiatric/ psychotherapeutic practice? Why (not)? If some adaptions are being made, regulations change, the technology improves, etc.., do you think it would make the use of AEI within psychiatric/ psychotherapeutic practice more/less desirable/practical/appropriate? Do you think the use of AEI is more appropriate in some areas of psychiatric/ psychotherapeutic practice than in others? If yes, which and why?

Are there any comments you would like to share? Or do you have any suggestions relating to this research?

Thank you for the interview. If interested, I could share the research. Hoping to finalize it around August.

Interview-Leitfaden Deutsch

Erläuterung der Ziele und Zwecke der Forschung

Künstliche Intelligenz ist derzeit ein sehr wichtiges Thema. Ihr Einsatz und ihre Anwendung haben in fast alle Bereiche Einzug gehalten, von der Wirtschaft bis zur Freizeitgestaltung. Es scheint so, als ob man im täglichen Leben nicht mehr um den Kontakt mit KI herumkommt. Die Forschung geht rasant weiter, und die Möglichkeiten der KI scheinen endlos und jenseits der Vorstellungskraft zu sein. KI wird häufig auf der Grundlage menschlicher Prinzipien und mit dem Ziel entwickelt, in bestimmten Aufgaben mindestens so gut wie der Mensch zu werden, wenn nicht sogar besser als er. Daher ist es nicht verwunderlich, dass auch im Bereich der Emotionen geforscht wird. Künstliche emotionale Intelligenz ist eine der Visionen für die Zukunft. Ziel meiner Arbeit ist es, herauszufinden, inwieweit KI in der Lage sein wird, emotional intelligent zu werden und welche Auswirkungen eine emotional intelligente KI haben könnte. Ich untersuche die Implikationen, insbesondere im Hinblick auf den Einsatz von Künstlicher (Emotionaler) Intelligenz in der Psychiatrie/Psychotherapie.

Ich möchte die Vorteile und Herausforderungen untersuchen, die EKI für diese Disziplin mit sich bringt und wie EKI die Praxis und das Verständnis dieser Disziplin verändern wird. Durch die Durchführung von Experteninterviews in diesem Bereich hoffe ich, Einblicke in die spezifischen praktischen Erfahrungen und Auswirkungen des Einsatzes von EKI in der Psychiatrie und Psychotherapie zu gewinnen.

Einholung der Zustimmung zur Aufzeichnung des Gesprächs. Die Gespräche werden an einem sicheren Ort gespeichert.

Das Gespräch dauert etwa 45 Minuten.

Themen	Interviewfragen
Allgemeine Informationen und Arbeitsumfeld	Würden Sie sich zunächst kurz vorstellen und erläutern was Sie mit EKI und Psychiatrie/Psychotherapie verbindet?

Allgemeine EKI (Anwendungen)	Was wissen Sie über den aktuellen Stand der EKI und wie sie in der psychiatrischen/psychotherapeutischen Praxis eingesetzt wird? Mögliche Folgefragen: Welche konkreten Anwendungen kennen Sie? Bitte erzählen Sie mir mehr über Anwendung xy. Verwenden Sie EKI-Anwendungen/Technologien in Ihrer täglichen Praxis? - Wenn nein, würden Sie es gerne tun? Wenn nein, warum nicht? - Wenn ja, welche und warum? Sind Sie an der Entwicklung/Forschung oder Erprobung von EKI-Technologien/Anwendungen beteiligt? Wenn ja, erzählen Sie mir bitte mehr darüber.
Chancen und Herausforderungen	Nachdem wir nun über einige bestehende EKI-Technologien/Anwendungen gesprochen haben, lassen Sie uns ein wenig darüber sprechen, wo Sie das größte Potenzial sehen, das EKI-Technologien der psychiatrischen/psychotherapeutischen Praxis haben könnten. Was denken Sie, in welchen Bereichen oder für welche Zwecke wird die KI von Nutzen sein und warum?

	Natürlich gibt es nicht nur positive Ansichten und Visionen zu diesem Thema. Wo sehen Sie die Herausforderungen, die mit der Einführung der EKI in die psychiatrischen/psychotherapeutischen Praxis einhergehen? Welche weiteren Herausforderungen (neben ethischen/technischen) sehen/erwarten Sie? Welche Visionen haben Sie für Ihre/ die zukünftige Forschung zu EKI in der psychiatrischen/ psychotherapeutischen Praxis? Wo glauben Sie, dass Sie auf Schwierigkeiten/Herausforderungen stoßen könnten und warum?
Erstrebenswert oder nicht?	Glauben Sie, dass die EKI ein wünschenswertes Instrument für die psychiatrische/psychotherapeutische Praxis ist? Warum (nicht)? Wenn einige Anpassungen vorgenommen werden, wenn sich die Vorschriften ändern, wenn sich die Technologie verbessert usw., glauben Sie, dass dies den Einsatz der EKI in der psychiatrischen/psychotherapeutischen Praxis mehr/weniger wünschenswert/praktisch/angemessen machen würde? Glauben Sie, dass der Einsatz von EKI in einigen Bereichen der psychiatrischen/psychotherapeutischen

	Praxis sinnvoller ist als in anderen? Wenn ja, welche und warum?

Haben Sie Anmerkungen, die Sie mir mitteilen möchten? Oder haben Sie irgendwelche Vorschläge zu dieser Untersuchung?

Ich danke Ihnen für das Gespräch. Wenn Sie Interesse haben, kann ich Ihnen die Forschungsergebnisse mitteilen. Ich hoffe, dass ich sie im August fertigstellen kann.

Appendix B

Interview Transcripts

Interview Dr. Ed de Bruin

00:00:03

Me: Yeah. Um, would you maybe start with telling me more about yourself? Like, what is your profession? What are you working with and what is your connection to the topic of AI and psychiatry?

00:00:18

Ed de Bruin: Yes. Um, I am an assistant professor at the Department of Psychology, Health and Technology, and my specialities in terms of research is sleep. Behavioural sleep interventions for children and adolescents mainly. But as a sleep scientist I need to be up to date on all current topics in sleep. So, over the whole population, which is a very broad field. And the second area of interest and focus of me is mindfulness. I'm also a mindfulness trainer and um, in terms of what my involvement is in, in the bridge between technology and psychology, I'm a psychologist. I'm not a psychiatrist. You mentioned the field of psychiatry. There is, of course, a large overlap, but the field of psychiatry approaches mental health and psychological problems from a medical field, whereas psychology approaches it from a behavioural field mostly. And my... I am involved in some courses that we teach into students concerning, for instance, a course that's called compassionate technology, in which we teach students about the use of technology in supporting mental health and how that technology can be applied in a compassionate way or even be compassionate. And another area where I'm involved in this research is in how we measure with the technology that we have well being. And that has to do with what is the best measure for well-being, what is well-being, how is it experienced? Are there physiological measures? Are there psychological measures? And and I am involved... My own research mainly was I'm currently am switching a little bit the field, but was applying online therapy for adolescents with sleep problems. So that is more or less the description of how I'm involved in this field.

00:02:55

Me: Mm. Very interesting. Very broad and very many topics. Like. Yeah. Um. Probably we could continue, like with specific applications you know, or. Yeah. Which are used or under development that are involving artificial intelligence or are linked to it in like this therapy context.

00:03:19

Ed de Bruin: Yeah. And there are two things that come to mind straight away that is automated chatbots and there is a long history of automated chatbots dating back decades that already there was research in. If you can automate, automate the responses of a chatbot, what happens, do people accept it. Um. And of course, with the current developments, this is going this is going more and more into a flexible application, whereas it used to be more or less pre-programmed responses. Um, and it showed from research in that time that simply having a chatbot respond as if it was listening to you. So for instance, humming, right? One of the mean, the conversational techniques in psychology, which you must be must, must have encountered as well. For instance, parroting, right. Responding, summarising what a person has said. So you mentioned this and this. Can you tell me more about that or. Mhm. Um, feeling lonely simply that. Right. Parroting what a person. So if you program those responses in the chatbot what would happen? And it appeared that people were very engaged with it even got somewhat addicted to continuing their conversation which was not really conversation but they experienced it as something that that helped him, helped him along. And you could speculate on what was happening. A part of what may have been happening is that simply having a sort of reflection, putting things out there as a narrative, already makes a person more aware of what is going on inside them. Whereas if you keep it inside, you simply it's something undefinable and untouchable and it's vague and you just simply experience how it makes you feel. Whereas when you put it out there, you might work through it and it becomes more clear to you and you might have clearer ideas on what to do about it. So there are some general factors. That with with even without, let's say, intelligence. But only the artificial part in responding is already enough for people to, to, Yeah, to, to have some effect. Mhm. But nowadays there are chatbots that go much further and yeah, you can have a complete conversation with it. And so far the idea has been to sort of use those chatbots as an add-on in therapy. So it could provide some psychoeducation. Um, it could provide perhaps some reflection for a person, but it seems that this might develop even further. The, the trend is that, um, therapists themselves are very conservative, like in any population to adopt new technologies. And as we all do, we feel that somehow a part of our work is unique, uniquely human, and we need to be there to do that. Mhm. Um that may be the case, but the question is what exactly is it that is unique and human to us that a machine can't emulate? And, and if we really look closely at that question, the idea that some that the machine may take over our job it vanishes because we still need to work it's just a shift of the place where we work. Mhm. For instance, developing new therapies, developing new ways to engage with the person with new approaches. Right. Not, not so much maybe the cognitive approach or the behavioural approach, but maybe a more mindfulness approach or an commit, a commitment, acceptance and commitment approach. And those are constantly under development and so that it shifts a little bit to work. So then the chatbots and technology may be able to apply some and do some of that work. But the development of the, the way the techniques, etcetera is still largely a human endeavour. So that is chatbots. Another point where it might be applied in current um, treatments is in, for instance, just-in-time adaptive interventions and those interventions. It appeared that if you have a therapy, then you have a lot of techniques. And when you are in a face to face situation, the therapist decides when and how and how intensely to apply a technique. For instance, let's say a behavioural experiment, a person may be afraid of large spaces, and so you do a an exposure experiment in which step by step the client goes through exposing themselves to that feared circumstances. And of course, this is a gradual exposure and. When to decide that and when to apply that. That is something that you can decide as a therapist. But if you have an automated intervention, it used to be that there was simply a sort of pre-programmed, step wise approach to this based on what we know that generally to apply this at this stage is best. But. If you can tailor that as much as possible to the client's wishes and needs, then it's more effective. So if at a certain point they become anxious about something, they can be helped by a by the intervention. Um, immediately adapting in what technique will be applied at that time. So it's an adaptive just in time, adaptive interventions. And there is a lot going on in that area. Um, which is also very interesting. Um, and it, of course there is a blend with, um, with chatbots in there and then a blend with some therapy guidance. And it appears that overall. From research. It appears that people there is if people have an online intervention, an automated online intervention. Um, it generally does have some effect, but as soon as they have the impression that there is someone behind the intervention actually looking at their material and responding or doing something, it boosts the effectiveness of the intervention. There is really like a boost in the effectiveness and it appears from research that so that it's important to have at least some therapist input in an intervention. But then from other research, it appears that it doesn't really matter who is giving that input, whether it's a specialist, whether it's an amateur, as long as there is the impression that there is something, someone giving some feedback. So again, it's a little bit unclear what it is that that makes it a fact. Is it simply having the, you know, being more committed because you may have the feeling that someone is expecting you to do certain things as a client? You're thinking, I need to show up or I need to do it right. And whereas if you're interacting with a machine that might be less prevalent. So that is a very interesting part of the, the j'étais and the combination with chatbots and the the interaction with a therapist.

00:12:09

Me: Mm hmm. Yeah, it's really interesting. Like the reactions you get from from the patients or clients. Um. That they actually don't care, like who's behind it?

00:12:20

Ed de Bruin: Yes, it's very interesting. It's the simply the fact that there is a human behind it or even the impression that there is a human behind it. Because how sure can you still be that a human is writing their response to you nowadays, Right? Yeah. And does that slowly, slowly, sort of fade away? Fade out, right. That slowly there's a transition and that people at a certain point, they don't care anymore if it's a chatbot or whether it's a human responding to their questions and their, uh, their concerns.

00:13:01

Me: Yeah, right. Um, you said that like, therapists are pretty conservative currently. Do you know if some of those technologies are already in use, or are they just like they.

00:13:14

Ed de Bruin: Are in use? But it's very in general in in mental health care, it's it's quite difficult to, um. Implement something. And that's not really different from any other technologies. Implementing other technologies is always a first adopters group that are very enthusiastic about something. And then there is a large group that slowly, bit by bit, transitions to adopting that technology. And then there is a group of late comers who resist, maybe even consciously or deliberately adopting a new technology that's very common and generic in in new technologies. But it appears that. Because mental health care has been organised, of course very, very, very thoroughly and there is a lot of money involved because you know, the insurance companies and so also concerned with this, there is a way of working that needs to be accountable. People need to be accountable for what they do to who and why and how much money it involves. And that makes it that is one of the factors that is a little bit resistant to change because a new therapy, you know, is it effective? What will it do? How much does it cost? And is there a person who is adept in that new technology, new intervention? And and then, of course, there is the factor of. Um, you know, the human behind it that is used to a certain way of working that they're very well, very well trained at and also very good at. Maybe they have years of experience. And so you don't want to, you know, don't change a winning team if you're very if you have good results with a certain way of working, you don't want to change it. But then again, so that's those factors all play a role. And then finally, there is the factor of, you know, simply your your job. You may have the feeling that your job is slowly being phased out by technology. But another finally. Another final factor that seems to play a large role is the consistency and the maintenance. There is a very enthusiastic group that does some research at the university. They do a new technology. They research it. They publish their papers. You know, they have 1 or 2 PhDs that do their promotion. They do their their PhD on that subject. And then after five, six years, it's over and the money runs out then and then? Then what happens? They have a very nice new technology. Who is maintaining it there? Maybe it's a software and you know, there is updates to the software and the software is not is not up to date anymore. Who, who and how is that maintained and who adopts it. And that is very difficult because the people who decide about these things. They you know, they say where the money goes and they have the same kind of limitations as what we've been discussing just before about their their knowledge about new technologies. ET cetera. ET cetera. ET cetera. Mhm. So there is a lot of initiatives, a lot of research that shows that there is a lot of possibilities and reality is very, very slowly lagging behind and slowly. And maybe there is a good thing in that as well because it means that any technologies that somehow don't that seem promising that may have some results but still are not too sure. It also shift shifts through the, you know, the the chaff right from the weed. It takes the ones that work and that really are solid and steady. They will stay there in ten years time. They're still available whereas the others, they're, they're gone.

00:17:34

Me: Mhm.

00:17:35

Me: Yeah, you're right. You already mentioned it's like a very visionary field. And that's what I also found. Like in my research, I was expecting to, um. Yeah, like compare it to the lot of research that is done actually. Also like in the Netherlands, I'm from Germany and here they are like way more conservative even and maybe way more hesitant to do it. So they do more like, um, fundamental research, but like not the application research and more like the clinical...clinical mental health direction because it's way more difficult to even get funding and the approval for doing those studies, those field studies with like real patients or clients. Um, so do you see like. Where are the most challenging parts about it like about all this AI applications in in this context? Um, what is most challenging about it?

00:18:38

Ed de Bruin: Well, there are a few challenges that we've discussed, a few challenges like the technology and the human adoption of technology. But another large challenge is privacy. Um, it's data and it's stored somewhere and can always be retrieved. And how do you deal with privacy if that if this application runs on a certain server in America? Um, do the privacy laws from Europe still apply? You know, they don't, you know, they're already so but there is a shady area there and, and as long as people are a little bit hesitant about it, can you use Zoom for your client? You know, your, your, your, your distant, uh, client contacts.

00:19:26
Me: Um.

00:19:27
Ed de Bruin: Those, those things are very important and also quite sensitive. So that is definitely a challenge. Also, how is the data being used if it is stored? Very often this data is stored and you get a disclaimer, uh, you know, informed consent that the data may be used to train the AI further. So what does that mean? Are people actually gathering your information, knowing things about you? And I can very quickly come up with, um, with maybe, you know, directions of your, um, diagnosis. So privacy is definitely an issue. Um. But I think there is also a very closely related to that. There is a an upside as soon as a person and especially in a medical field, if you ask a chatbot. You type in your your symptoms about maybe a physical condition. You know, you have a stomach pain. You have it for some so long you've also had some pain in your foot on your right foot, and you have a headache when you get up. And you write something about your lifestyle and it immediately gives you a diagnosis that they can't come up with in a hospital as soon as that happens. People are going to say, I want the AI because it's a life or death situation and they don't care very much anymore about privacy, especially if it's, you know, if it's really very effective and very fast. Right. So I think that's another challenge that is also an opportunity in the sense that if we somehow know how to harness and mobilise the potential. It can be do wonders. The knowledge, the concentration of knowledge and the application of knowledge is very accessible for anyone. And I think that can help the medical field. Also the mental health field very, very, very much.

00:21:52
Me: Yeah that's right. Yeah. You already started to talk about some possibilities that are there. Um, where do you see the, the most promising directions of research that are also, like, more close to being really implemented?

00:22:07
Ed de Bruin: Yeah, I think, um, we have a system of classification of mental disorders which exists for about 40 years now, and they are continuously under discussion. Whether, you know, whether what is the validity and there is a sort of self maintaining self-fulfilling prophecy. If

you have a classification, then there is for researchers a sort of there is an importance of, you know, designing their research according to those classifications. So, for instance, insomnia has certain criteria whether a person has insomnia or not. And so you're going to look for your clients if they have insomnia and then treat them. But you're not going to look for what is their problem. You're simply looking for you through a lens is your classification. And if you simply pile up all that data in large, large, large quantities and do research with those large quantities of data, look for patterns you might come up with an alternative classification. It might be much more applicable, it might even be a classification that is somehow adaptive to cultures, might be adaptive to age, to gender, to, you know, to sex. It might be adaptive to individuals even and developmental issues. And then you can come up with treatments, with diagnoses and treatments which are much closer to what a person, an individual needs. And I think that is something that we're. Slowly, slowly. I mean, this is something that might take another 20 years or so, but I think we're going there slowly. Yes.

00:24:17

Me: Yeah that's really interesting. Also, I feel like there's this paradigm shift from this classification systems to become like more loose and less strict and more individualised. And also, yeah, probably AI could help in that and see patterns humans don't really see or yeah, kind of like are hesitant to to see because they are so stuck in their frames and I mean they learn them during their whole education. So yes and.

00:24:47

Ed de Bruin: So is it true and that's that's a little bit a generational thing, right? I mean sometimes it takes a generation to shift because the next generation is more open to those new possibilities and they bring it into the work field. And so the shift is coming with the people that so slowly shift. So I think that's, uh that takes time.

00:25:17

Me: There is actually an like online symposium tomorrow about this. Um. Reviewing critically this diagnostic practices. Um, yeah. So I think it's really interesting that all this changes and also like what role technology might play in this change. Yeah.

00:25:38

Ed de Bruin: Yeah, I think that I just one thing I think that technology ultimately it will be of a great help.

00:25:46

Me: Mhm.

00:25:47

Ed de Bruin: Um, but yeah, we, we might need to give something up in exchange for that. I don't know what yet but maybe privacy, maybe privacy is a, is a, is a concept that is not that important actually. But anyway. Yeah.

00:26:06

Me: Yeah, at least to some extent. And yeah, within the framework of helping other people.

00:26:12

Ed de Bruin: Exactly. Exactly.

00:26:14

Me: Yeah, right. Um, do you see, like, any other possibilities where I could find its application in this Um, context?

00:26:25

Ed de Bruin: So I was speaking about diagnostic, sort of the diagnostic direction. Um. The the possibilities of AI are, of course, sort of synthesising large quantities of what is already known generally. Right. Even generating responses to questions or, you know, generating images. That's based always on what is already out there. So it's like a sort of synthesis. So it depends very much on what what we ask AI But. I think that's. I think the first the most important one is the classification, the diagnostic part. I mean, of course the helping in responding in treatment, etcetera. But I don't know. That's hard to say. What is hard to say is what is important to humans. Is it you, other human contact, contact or is it interaction? Any kind of I mean, you you might say that there is a residue of a human in the response of AI because it's all based on previous human. You know, generated information. Right. It's information is so there is still humanity in the response of AI, even though it's artificial intelligence. Um. But it's hard to say how how whether that's whether that really goes all the way. That it's that it doesn't matter anymore.

00:28:09

Me: Yeah, right. It's pretty unpredictable. Like, yeah, the research is going on so fast and people are. Yeah. Not yet at the stage of fully accepting it. No. Um, also, do you know, like, what patients think about this? Technologies that are already used, Are they, like, more open than the conservative therapists or are they usually.

00:28:34

Ed de Bruin: Yeah, they usually they are more open than conservative therapists in the sense that I mean very much in a similar way as the therapists, but they have one more motivating factor and that is they want to be helped. So if there is any indication that a technology can offer some form of help or support, people are very willing to adopt it. Mm.

00:29:06

Me: Mm.

00:29:10

Me: It is interesting because when you usually talk to people, they are like, Oh, I don't know. It's like so sensitive topics and yeah, but. I've never really spoken to like patients who are using technologies. I know that in Germany we have like some apps you could prescribe via the insurance for your clients. Now, um, I don't know how it is for the Netherlands, but at least there are some options now you could use or try.

00:29:42

Ed de Bruin: Well, I'll give you one more example of how that works, right? So there are people with, for instance, narcolepsy, which is to give it to come back to my field. They have a sleeping disease and narcolepsy is. Um, you cannot cure it and you cannot completely adequately medicate it. You cannot completely provide medication that completely prevents it happening. So usually one of the important treatments is lifestyle advice, right? People need to avoid certain foods and drinks and they need to, you know, rest enough, sleep enough. ET cetera. ET cetera. ET cetera. But there is one area of research that's already ten years old or so, so that people who have narcolepsy there is a shift in their body temperature. The, um, the, what's it called? The proportion or not the proportion, but the balance between the core body temperature and distal body temperature. So the, the hand and so on and how, and the balance between those two. And it shifts. There is a noticeable shift that predicts. A couple of minutes before it happens. A narcolepsy attack. So that means that those people could be helped if they are, for instance. You know what to do when you have narcolepsy and you want to go in your car somewhere alone, right? Yeah. Can you do that? So that that technology might, if you have something that you know that you attach like a sleeve or a ring, you know, and and and you simply have an app warning you, you know, you need to take this or you need to quiet down or you need to sit down. 5 minutes or 10 minutes before it happens. That's a great that's a great help. And I'm sure that people with narcolepsy would immediately adopt that. Immediately, without question. And there's a lot of those kind of things that we know from research already and that are waiting to be applied. And then, of course, the question remains, why is it not applied yet? Well, there are several things. Again, the privacy and the but mainly the funding. You know there is this large, um. A business of apnoea devices, right? People that have apnoea breathing problem while they're sleeping and so they miss a part of their sleeping architecture and it's very deliberate. It's very. What's it disabling if you have that cognitively and the ageing and so on. So there's a large technology business, large, I mean billions of, of of dollars yearly that go into that business. And Philips is a large manufacturer of those devices. A couple of years ago it appeared that the masks that they produced, they had rubber on the outside covering the face that appeared to be a little bit poisonous. And so they had to take back all those masks out of the market. And it was $1 billion billions of dollars problem for them. So it's that's very, very much money. And that's why that industry is evolving very quickly. You know, the adaptability of how that responds to your breathing, which is another again, you know, AI technology could be implemented there as well. How it adapts to your breathing and when it adapts. Right. How much pressure it gives in the air. But anyway, there is a lot of incentive for technology companies to develop that. And that's something that we yeah that money plays unfortunately plays a large role in in the in the implementation of those technologies. But what I said like this kind of technology such as with the narcolepsy and there's more of those technologies, people with epilepsy, I'm sure there are some things that you could do with that as well. Um. And then all the technologies mean the quantifiable self. You know how we monitor our health and how we monitor our behaviour according to our health? There's yeah, so that's, those are definitely areas that we'll see a lot of development over the coming years. Mm.

00:34:40

Me: Yeah, I agree. Especially those like warning systems might be very helpful and very useful and patients would really want that because there's no not many downsides. I would say.

00:34:52

Ed de Bruin: There's not many downsides and it may improve the quality of your life immensely. Right? How independent and free you are in deciding where you want to go and what you want to do in a day and when. ET cetera. It's immensely. Yeah.

00:35:08

Me: Yeah, I agree. So, like, what I hear from your answers is that you're mainly, um, in favour of those technologies and think they are like a desirable add on for therapy. Um, there's this, um, regulation of like, the human in the loop. Do you agree with that? That there always should be like a human involved in this process? Or do you think like technology at some point could also act as an automated agent?

00:35:41

Ed de Bruin: Um, I think it is not desirable to do that technology as an automated agent. It depends on what area. But I do agree with the human in the loop. Of course you come very I mean, you come to very philosophical questions like what is the purpose of life? Yeah, right. If the if the the purpose of life for humans, at least it's a question for what is the purpose of human life for or for me even. And then you're asking, so how much agency do I have, how much decision room and where and how and. I think that is something that overall people usually take the the interference of technology to too anxiously, perhaps a little bit. But. Um. I think we we would come to a standstill if AI is applied completely independently because it is based on what is already there and I'm not sure how capable it is of. Um, purposely developing into something new. Evolving, so to speak, and whether and how that corresponds with human evolution. But I think we come to very broad questions there and. If you look back, I mean, there is always the idea that we're into something new that is very unique and very for this time that has never been done before. I mean, on the one hand that's of course true, but on the other hand, in terms of human experience, is it true? We we we were born we we live a life with all kinds of experience, and then we die in that journey. There is nothing new. So so I in that sense, it's. It's another aspect, I mean, in terms of the human experience. So yeah, so I agree with that that loop. Human in the loop, but I think it is a little bit dependent on where exactly.

00:38:07

Me: Yeah.

00:38:07

Ed de Bruin: What we're looking at there is this also I mean that's a very old conundrum already in psychology. There is. And you know this probably right, because you did conflict risk and safety. And you know, this thing that one of the main causes of human accidents in in in big. Operations, for instance, there is this whole system set up to warn you about potential life danger in in when in electricity plants. Right. So you shouldn't when there is an alarm going off, you shouldn't open the gate to come close to the whatever it is, the machine. And strangely enough, people tend to switch off the alarm because it's annoying them and then still go in to have a look what is happening and then an accident happens. Yeah. So we, we set up all those procedures to be safe and then we sort of break the rules.

00:39:13

Ed de Bruin: And.

00:39:13

Ed de Bruin: We still think that's a very psychological. Um. Yeah, very peculiar human thing. I think it's interesting that I mean, because we were talking about which area, where do we apply this intelligence? Um, yeah. We are driven to experiment. So.

00:39:39

Me: Yeah, right. Yeah, I think we discussed that also. Like with little children that always do what they are not allowed to do.

00:39:46

Ed de Bruin: Exactly. Exactly. We are playful. We're always looking at at at borders and at limits.

00:39:52

Me: Mhm.

00:39:56

Me: That's interesting. Yeah. Nice perspective. I haven't thought about that, actually, but it fits nicely. Um, you also said that you were developing or. Yeah, like, previously, um, some kind of online therapy for those. Um, was it insomnia or sleep patients? Insomnia.

00:40:17

Ed de Bruin: Insomnia.

00:40:18

Ed de Bruin: Yeah.

00:40:19

Me: Can you tell me a bit more about that?

00:40:21

Ed de Bruin: Yeah. We treated adolescents in groups, group therapy, face to face, and in individualised internet therapy with a therapist with some chat and some interaction and some personalised advice and and then with some automated exercises based on a protocol that's shown to be effective. And we researched the effectiveness of it to establish whether it's effective that treatment form for adolescents as well and how internet and group therapy, how they compare to each other. And it appeared that effectiveness in both forms of treatment were very much similar, um, and with a slightly less commitment to therapy in the internet therapy, um, but not significantly less. It was only a trend and um, yeah that's that's that's more or less in a nutshell it.

00:41:30

Me: Mm That's nice. Yeah. So it seems like people are actually accepting it and are willing to use it also in this area.

00:41:40

Ed de Bruin: Yes that's.

00:41:41

Ed de Bruin: Yes, people are willing to also adolescents are willing to accept it in that area. And um, and to use the, the, the online intervention options that that they have. Yeah.

00:41:55

Me: Did you also do like some kind of evaluation afterwards, like how they liked it or like some acceptance by research.

00:42:03

Ed de Bruin: Yeah. In the pilot study we did and. There was a there was. A complete or complete. There was a very high level of acceptance of the technology that we applied. Um. The main thing was that the people had the impression that they had a specialist dedicated to their issue. Because sleep is so general in that sense. Maybe it's it's for all kinds of conditions that it's important that you have a specialist. But in sleep, there is so much knowledge out there that is a little bit murky. It's not completely clear how well based it is. So they very much find a high importance in a specialist being committed to that. And that relates very strongly back to that sense of having someone, even if it's online therapy automated, and you still have the impression that there's someone looking at your data that's important for people. Yeah.

00:43:13

Me: Yeah. Nice. Very interesting. Interesting. Um, I think that basically was it from my side. Do you have anything to add or some things we did not discuss yet?

00:43:25

Ed de Bruin: No, not for the moment. I don't think so.

00:43:30

Ed de Bruin: I don't think so. It's an interesting topic. Definitely.

00:43:34

Ed de Bruin: Yeah. No, I don't have any I don't have anything else to add.

00:43:38

Me: Okay, great.

00:43:39

Me: Um, you already mentioned, like BMS lab. Do you know any other people I could probably approach?

00:43:46

Ed de Bruin: Uh, you might approach also.

00:43:49

Ed de Bruin: She's very closely. Maybe she doesn't. She probably tells you something new, but she's doing a promotion. Research, PhD research. Charlotta von Lothringen. She's also at UT and is being guided by Matthias Norzi, who is one of her supervisors. And she's very much into also that compassionate technology and. And she might. You might have some some insights. That is one I think BMS lab would be good, especially if you would talk to some of the programmers from their perspective. That might be interesting. Um. Other than that, nothing comes to mind straight away, so.

00:44:35

Me: Okay.

00:44:35

Me: But thank you. Yeah that already helps. Also, thank you for your time. It was really nice talking to you. Very nice insights.

00:44:43

Ed de Bruin: Well, you're welcome. And I would say good luck. All the best of luck with the continuation of your work. And I'd be interested when you're finished to to receive your your thesis.

00:44:57

Me: Yeah, sure. I can send it.

00:44:58

Ed de Bruin: It would be great.

00:44:59

Me: Yeah. I hope I will finish around August, early September.

00:45:05

Ed de Bruin: Great. Well, good luck. And. Yeah. Thank you. Have a nice day.

00:45:09

Me: Yeah, Same for you. Bye bye.

Interview Caroline Bollen

00:00:24

Me: Could we start maybe by, um. Would you introduce yourself and. Yeah. What you do in your work and what links you to the topic we're discussing today, like emotional AI and probably also psychiatry if you have some experience there.

00:00:42

Caroline Bollen: Yeah, sure. Um, so I'm a PhD candidate at Delft and my dissertation is on empathy, communication technologies and neurodiversity. Um, I'm mainly focusing on building a new conceptualisation of empathy that is more inclusive towards autistic empathetic experiences and that we can also use to better evaluate communication technologies and the ways in which they can like facilitate empathy or stand in the way of it. Um. And so for that, I've not been explicitly working on whether I can be emotionally intelligent and those use cases. But yeah, I hope I can still think along with you because I have been working on emotions and technology and also psychiatry in the sense of neurodiversity and autism. And I also have a background in neuroscience.

00:01:40

Me: Oh, nice. Yeah, I think like autism and neurodiversity are some of the aspects which are often mentioned to be covered by AI technologies more easily. I don't know, but at least that it could be helpful because you can use like avatars or something to train or for the autistic people to learn. Yeah.

00:02:04

Caroline Bollen: Yeah. Some time ago I also did an interview like this with another student team who are working on, like, emotionally intelligent robots that can help autistic kids learn skills.

00:02:22

Me: Great. Um. So, yeah. What do you know like about the current state of artificial emotional intelligence and how far it is used? Or we discussed already that it's not really used in practice, but probably in research.

00:02:42

Caroline Bollen: Yeah, I'm not really up to date with the state of the art developments in that field.

00:02:47

Me: Okay that's totally fine. Um, but you already talked about, like, specific applications. Um, do you know some more applications or, um, just the app (Voidpet) you talked about (before recording started)?

00:03:06

Caroline Bollen: Um. Yeah. And other like there's like so many applications of large language models at the moment. You know, like GPT that also kind of try to mimic mimic at least some emotional intelligence and um, and you can ask them to respond in a specific emotional style and things like that. They're quite accurate sometimes in, in actually mimicking that kind of thing. Um, but not in, in psychiatry.

00:03:38

Me: Okay, Well, there. Are people who are like trying to test out if you ask mental health questions and how we can respond also compassionately. Well. Mhm. Yeah. Also read about like people using ChatGPT as their personal psychiatrist.

00:03:56

Caroline Bollen: Yeah. Yeah.

00:03:58

Me: So that's pretty weird. But, um, what technologies are you working with on like it was information technologies, right?

00:04:07

Caroline Bollen: Communication Technology.

00:04:09

Caroline Bollen: Communication. Yeah.

00:04:10

Caroline Bollen: Um, and I've a specific focus on alternative and augmentative communication technologies. So communication technologies for people who, um, are non-vocal or minimally vocal. So they, yeah, use other technologies for communication. But I've also been working a little bit on social media and other more widespread communication technologies.

00:04:35

Me: Mhm.

00:04:37

Me: Um, and you're working on it like just with the focus on the two groups you said, like autistic people and neurodiverse people.

00:04:48

Caroline Bollen: Yeah, I have. There is a specific focus on that specifically because they are usually left out of the conversations about empathy. Um, but I've also moved on to more like the broader population. Um, but like as a, as a case, it's a very interesting one.

00:05:07

Me: Mhm.

00:05:07

Me: Would you like to tell me some more about it, especially like the empathy component.

00:05:14

Caroline Bollen: The empathy components of?

00:05:17

Me: Within the technology is like, what is your focus from which side are you approaching it?

00:05:24

Caroline Bollen: Um, yeah, I. So I first started with like the question of what empathy is. And I quickly figured out that there's no agreement or consensus at all. And there are also some problems in the field itself on how to measure it and how to quantify it as people try to try to do and the problems in there with excluding autistic empathetic experiences. Um, so that's why I've instead chosen a virtue, ethical approach to empathy and conceptualised empathy as a virtue, and specifically one that, um, helps you attend to differences and similarities in experiences. So, um. You know, there are some things in our experience that we share. We're both human. We're both women, for example, but also, like you are not me. And it's kind of finding that balance between identifying with the other, but also knowing how to take a e perspective and knowing what is different. And you can't understand.

00:06:28

Me: Mm. Yeah. That's very interesting. Um, sorry. Um, so you, um. You said you're more developing, like, a theoretical framework about it.

00:06:45

Caroline Bollen: Mhm.

00:06:46

Me: Um, are you applying it also to, like, um, practical technologies then, or just more to the empathy part of it?

00:06:57

Caroline Bollen: Um, yeah. I'm also applying it to the development of communication technologies and these technologies in specific, but also more general. But that also stays on the level of theoretical framework and theoretical guidance. I'm not actually collaborating with any engineers or designers at the moment. Um. But I have been working on how you can apply this theory onto the development of technologies.

00:07:28

Me: Oh, great. Where do you see, like, the most promising possibilities for this approach, are also the technologies to come, probably.

00:07:43

Caroline Bollen: Um, well, what I really hope is that, um, technologies, or at least in technology development, we better integrate natural tendencies that are not very desirable. So, for example, regarding empathy, we have as human beings a tendency to, to better empathise with people who are more like us and not people on the other side of the world or people of different gender, different ethnicity. And that is, of course, pretty problematic. So I hope that we can better like integrate what we know of of these undesirable human tendencies and and develop technologies that actually work along with the intention to do good rather than against it.

00:08:32

Me: Mm hmm.

00:08:33

Caroline Bollen: Yeah that's also kind of what I found. Like, as you said, there's like no consensus on definitions for empathy or I was looking more on emotions in general, but it's pretty similar. And the most yeah, critical point was actually like this lack of diversity within the theories and measurements they they developed, especially with AI. There are yeah, there's always the problem with bias and that the datasets are coming from like more western white countries. Yeah. Um, and especially they wrote a lot about like indigenous cultures where like the expression of emotions is so different that you cannot generalise it kind of. Um, so yeah, I think that's a very important point. Um, yeah. Do you see like also probably a challenge in it.

00:09:32

Caroline Bollen: Well, the challenge that you just discussed that we are like like perpetuating these narrow norms. Um, I'm also always with these kind of technologies worried about privacy and how they are being used and developed. Um, for example, there is this, uh, empathetic AI called replica that you can buy and that is kind of like it is marketed as a solution for loneliness. Um, which of course loneliness is a very big problem right now. And it's, there are like people who buy that and the people who they market it towards are in a vulnerable state already. And um, it's kind of a question of like, are you actually helping them or exploiting their vulnerability? Like that's even like beyond the question of whether an AI can actually substitute human relationships. Probably not, but it could maybe help you. Just like with therapy, it cannot, of course, replace a real therapist, but there's such a large barrier to getting therapy at the moment. It might help some things. But yeah, this, this bigger frame of, um, the system in which these technologies are being produced that's something I really worry about.

00:10:50

Me: Hmm.

00:10:51

Me: What do you mean? Like the system in which they are produced. Like the commercial sector? Yeah.

00:10:56

Caroline Bollen: With the. The commercialisation of it and the. The aim, of course, to make profit out of the vulnerabilities of of people.

00:11:06

Me: Mhm.

00:11:07

Me: Yeah that's actually what I also thought. Like most of those technologies are that are available are. Yeah. Produced by those big companies. Um, and they are like hide it as wellbeing apps or something, something that can be implemented now even though it's not like specifically in therapy, but it's like a stage before, I would say because it's open access, more or less. Some are paid but some are not. And the things which are actually like developed by researchers with like no financial interest in it. Um. Yeah. Which are backed up more in scientific evidence probably, and more thought through and more ethically, um, often don't make their way to the market, I would say.

00:11:56

Caroline Bollen: Yeah.

00:11:57

Me: So that can be really problematic, I guess.

00:12:00

Caroline Bollen: Yeah.

00:12:02

Me: Um, do you see any other challenges? Probably.

00:12:11

Caroline Bollen: Um, yeah, one specific one is that it will change how we interact with each other because I made like kind of a bold statement of like, they can never substitute real human relationships. But I am curious if in the long run people might have like different expectations of how a something responds to what you are saying. And um that so for example. So a chatbot for example can be endlessly patient if you like. Keep repeating yourself or something, but you can't expect the same thing of a human and that we then disvalue the things that are uniquely human. And uh, yeah, instead value some things that technology can bring that humans cannot. As always, as is already happening in so many other forms of automation, you know that we value things like speed and efficiency more than craftsmanship or creativity.

00:13:18

Me: Yeah that's right. Do you think that is especially an issue with like those emotional topics or like these empathy topics? Because it's even like a layer deeper, I would say, than craftsmanship or something. Like to. To what extent?

00:13:42

Caroline Bollen: Um.

00:13:43

Caroline Bollen: Yeah, to the extent that people are held back in developing the virtue of empathy, as I use a virtual approach to empathy, which then again stands in the way of building meaningful human relationships, which then builds a circle of like, it's better to interact with a technology than with a human, because you don't develop the skills and virtues needed to develop meaningful relationships. And well that's a very negative, vicious circle to end up in.

00:14:18

Me: Yeah. Like why is technology preventing people to develop those skills needed?

00:14:29

Caroline Bollen: Um.

00:14:31

Caroline Bollen: Well, if it. If. Well, this is like super hypothetical, right? And also, like in a negative lens, because I also think that we can, like, steer it in a way that it is more possible, but like in a more dystopian view. And if people would have less practice in actually engaging with other humans as a prevention for loneliness instead of like a easier fix with technology. Don't get the opportunity to engage with other humans and that is that is needed to build these kind of skills and virtues.

00:15:08

Me: Yeah, I think that's pretty much in line with what I thought because like often a point of critique is that those findings, even though they are like reliable in the laboratory setting, they cannot really be transferred to the real world. And people especially I think I read something about autistic people cannot really transfer it into the real world practice. And that can be pretty problematic. I mean, what would be the purpose then to pursue those approaches? Yeah.

00:15:42

Caroline Bollen: And if you then still do actually apply them to practice, then it can do more harm than good. Mm hmm.

00:15:50

Me: Yeah. Yeah that's right. Um. Okay. You already discussed that. Yeah, probably a bit too related to the question about the possibilities. Um, like we already discussed, it's, it's a very visionary field currently. Um, and it's not sure in which direction it will develop. Um, like what are your visions or desires for, for possible direction?

00:16:38

Caroline Bollen: Um.

00:16:40

Caroline Bollen: Yeah, my hope is that the human or potentially other animals, but at least living things stay central. And we um. And we view the. We view these technological developments as Aids to help us and um, with an acceptance of the. Of both the good and the bad about being human. So. So what? Previously mentioned about like productively working with bad human tendencies and actually centring. The vulnerabilities and strengths of being human. And in that way I can definitely see a role of these kind of technologies in developing certain skills or having some kind of support when there really is none but really centring. As long as it stays like in a kind of a second position to the human right.

00:17:46

Me: So like the, the human in the loop, which is often mentioned or even like that, the human is not only in the loop, but has the control over the technology.

00:17:57

Caroline Bollen: Yeah.

00:17:58

Me: Okay. Mhm.

00:18:00

Me: Um, and do you have like a specific example where those technologies could help, especially those. Yeah. Emotional technologies or if you can make them emotional at least.

00:18:14

Caroline Bollen: Mhm. Um.

00:18:20

Caroline Bollen: Well, I can imagine some, um. You know, some situations when you would. There are some things that like kind of are like an emotional response that could help someone in a specific situation. Um, for example, when you feel like you're panicking, for example, and there's really no one around to help you. Um, and that there is. A something that can help you to breathe and to calm down and say it's all going to be fine and this is going to pass and things like that. That is not, of course, actually a compassionate other human being. But if if this could prevent a panic attack from happening that's great. Right? But for that, you wouldn't even need to explain who you are. That can actually be a quite easy technology that can do that. Even just a recording of someone. Those things already exist, actually.

00:19:22

Me: Yeah that's right. Um. True. Like. Like an emergency help or whatever... support technology.

00:19:33

Caroline Bollen: Yeah, but that actually doesn't have to be necessarily be intelligent because there is a human who is intelligent and, um. Yeah. You have like all these, these, uh, wellbeing and meditation apps. Um that for example, the one I know best is headspace. I don't know if you know it, but it's a very large meditation app and there's like for every situation that you can possibly imagine, there is some kind of exercise and it is made by humans and also recorded by humans. And you as the user choose. Okay, now I am nervous about a flight, so then I can listen to this. But actually the app itself doesn't need to be intelligent at all. They're just different recordings. And you as the human are the intelligent one that picks the right one for the right situation.

00:20:25

Me: Mhm.

00:20:26

Me: But even if it doesn't have to be intelligent, it still has to be, um, empathic. Don't you think so?

00:20:38

Caroline Bollen: Um.

00:20:41

Caroline Bollen: Well, it's then it's still kind of a parasocial relationship, right? There's not actually another human who is listening to your story and what you are going on, what is going on in your life at that specific moment and your unique experience. So I'm very hesitant to ever call it actually empathetic.

00:21:03

Me: Yeah that's right.

00:21:05

Caroline Bollen: But more of a mimicking of an experience of being empathised with that can like very pragmatically have benefits in specific situations.

00:21:14

Me: Mhm.

00:21:15

Me: Yeah. I'm thinking about the situations you talk about like panic attacks or something and I think that at least um, the technology should give the user the impression of Yeah. Like kind of something that is similar to empathy. Um, yeah. To just ease the, the patient or the people approaching the system and yeah, as you said, it doesn't need to be intelligent, but it at least needs to be like. Somehow user oriented and empathetic.

00:21:54

Caroline Bollen: Yeah, but I think there the real empathy comes from the designers and engineers behind the technology. They need to be empathetic towards their users and future users, not the technology itself.

00:22:09

Me: So it's more like about the input that it's that it's created by humans.

00:22:16

Caroline Bollen: Created by and for humans.

00:22:19

Me: Yeah, I think that's also like the the common practice currently at least what I read about, um that those responses, especially in those therapy apps or chatbots, um, are at least like that, they're not really using generative, um, approaches, but more like those natural language processing and that are like pre-developed question-answer patterns which are approved by real psychologists. So exactly the system can just choose which to use when. Yeah. But like it cannot modify it or anything.

00:22:57

Caroline Bollen: Yeah.

00:22:59

Me: So because there.

00:23:00

Me: Are so much, so many risks.

00:23:03

Caroline Bollen: Yeah, right.

00:23:05

Me: So because they are like created by humans, it's still the empathy is coming from the human itself and not from the technology. You're right with that.

00:23:15

Me: Mhm.

00:23:25

Me: Yeah. Um. Probably like you're also already said it's more like a dystopian view you're you're pursuing to it. Um, do you see any, like, problems that might arise if those technologies become more prominent or more? Um. Yeah.

00:23:48

Me: Yeah.

00:23:49

Me: More integrated into society without any restrictions or whatever.

00:23:59

Caroline Bollen: Um.

00:24:00

Caroline Bollen: Yeah. Some sort of dependency relationship between the, again, the humans, not necessarily between the user and the technology, but the user and the people who make the technology and who own the technology.

00:24:15

Caroline Bollen: And the vulnerability.

00:24:16

Caroline Bollen: Of the user.

00:24:18

Caroline Bollen: And.

00:24:18

Me: How far like from the people developing it and the users?

00:24:23

Caroline Bollen: Yeah. Um, so once you've become. Because it is so, like, intimate and emotional. Um. There is some kind of like a dependency or vulnerability relationship that you can build towards using such an app or a chatbot or something. And that kind of puts the people who create the technology in a position of power of what they how they price the technology, what they do with the data they gather etcetera.

00:25:03

Me: Mhm.

00:25:03

Me: Yeah. No I get it. I just saw the, the line from the user to the technology or the developers but not the other way around. But this power relationship really makes sense.

00:25:14

Caroline Bollen: Yeah. Yeah.

00:25:15

Me: And it can be really problematic I think if there are no rules and they can just do whatever they want. Yeah.

00:25:23

Caroline Bollen: Um. Yeah.

00:25:26

Me: Also, like, um, what I read about was the dependency from the user about like, because technology is like available 24/seven and a real psychologist is not. So they. Yeah. Develop those dependencies very quickly because they rely for each and every problem they get in their life just from daily tasks or something. They always consult the apps or Chatbot or whatever they use, and that can be really problematic. I think there's not so much yeah, practical experience now, but there's a tendency at least observable.

00:26:06

Caroline Bollen: And that can also then stand in the way of the development of other skills like self sufficiency and solving problems by yourself and patience.

00:26:19

Me: Mhm. Yeah. Yeah.

00:26:22

Caroline Bollen: That might also influence what we expect from other humans and our relationships with other humans.

00:26:28

Me: Mhm.

00:26:29

Me: Um, you referred back to the skills you previously talked about like the skills you need for empathy. I haven't asked about it yet. Um, did you develop like your own set of skills which are needed for empathy or what other skills you suspect to be needed for being empathic?

00:26:50

Caroline Bollen: Um, well those what actually argue is that there's no specific set of skills, but that they are actually dynamic also because of technological changes. Um, so an example that I often use is that in, you know, old school, real life communication. Reading each other's facial expressions and understanding intonations are very important in like trying to understand what the other person is, is feeling and thinking. Um, but when you are texting with someone else, it's that those skills are not relevant at all. And then it's better to like being able to interpret text messages and maybe use emoticons and understanding them properly and know the like social norms about that where you can clearly, clearly see, for example, a difference in generations of how people use emoticons and what they mean with them, how they interpret them. Uh, so that's for example, something that has already changed in the skills needed for empathy in, in the current, uh, practice.

00:27:55

Me: Yeah that's actually a very interesting example because what I recognise is like that in written messages, there are often so much um, misunderstandings happening just because you're lacking the whole context and you don't have the person in front of you. Um, yeah. So I can relate to this example, actually.

00:28:17

Me: Yeah.

00:28:18

Caroline Bollen: Yeah. But also that's a different skill that you can develop more if you practice it more. And um, yeah, we'll relate it to autism. Reading facial expressions is often something they struggle with, so it actually changes the disadvantage that you have in compared to real life conversations where you do need to rely on facial expressions and actually having that removed kind of levels, the playing ground.

00:28:49

Me: Yeah.

00:28:50

Me: Nice. Nice comparison. Yeah. Um, do you think in general that, um, I. Or like something that could be called emotional AI, um, would be like, a desirable tool for therapeutic purposes?

00:29:14

Caroline Bollen: That's kind of a big question. Yeah. And I don't have a clear yes or no answer, obviously. I tend towards no just because I see so many pitfalls. But on the other hand, I don't think it would be wise to just say no and not explore the opportunities, particularly because there's such a big need for better mental health care, and there's really not enough practitioners and not enough funding. I still think it's better then to change the health care system in that way. But like in the meantime, or if we can really not make sure that it it it is just, I think, good practice to consider how we can use technology to do so. Um, it's just I would be very careful with considering AI to be actually emotional intelligence and to be very suspicious of what it can and can't do. And not lose focus on changing the health care system in the meantime. Definitely not an excuse.

00:30:24

Me: Yeah, I agree with that. Um.

00:30:31

Me: Like.

00:30:33

Me: If there would be some adaptations within the health care system, the general system, some regulations on the technology and the technologies probably also improve. Do you think it would make it more desirable or less desirable?

00:30:53

Caroline Bollen: Well, if you say it this way, obviously more desirable, but still, it's very yeah, it's very difficult to say if in such a hypothetical scenario, I think it should really be considered in a case by case basis of a very specific technology and the specific way it will be implemented and used and all those details that it really hangs on that.

00:31:19

Me: Yeah that's right. Um, do you think or probably you also have experience, like with the groups you're focusing on? Um, if they tend to rather accept the technology as a support system or something, or are they rather hesitant to use it?

00:31:40

Caroline Bollen: The technologies that I'm working with or the emotional technologies that you're referring to.

00:31:47

Me: Both. Um.

00:31:56

Caroline Bollen: Well, the thing is with the technology is that. I've been mainly focusing on. There's a lot of societal stigma involved with kind of completely changes the the landscape of what we're discussing, I think with a lot of, um, things that in the use for psychiatry, for example, there's actually probably more stigma on going to actual therapy than using such a technological tool that that. So it's kind of a opposite landscape there. Also an interesting angle, by the way, the stigma on mental health care. Um, yeah.

00:32:41

Me: Okay. Yeah, I think.

00:32:47

Me: That basically was it from my side. Do you have any comments you would like to share or any suggestions relating to my research? Probably.

00:32:58

Caroline Bollen: Yeah. I'm quite curious to hear what you've done so far and what you've found so far.

00:33:04

Me: Mhm.

00:33:05

Me: Um, yeah. Basically I just started with a very negative attitude towards it. Um, and I was also like, you could never apply it to the psychological sector because there's so many pitfalls and so many unsolved problems. And actually the more I read about it. I kind of follow the trap to to believe that it might be desirable. But yeah, I always have to remind myself about the negative sides because they often get neglected in those. Yeah. Euphoric articles about the great possibilities that might come along. And I think, yeah. They are actually some really interesting possibilities. And yeah, in the end, I would also say it's like it's still this mimicking of emotions and you're still manipulating people to a certain extent, especially when they are in such a vulnerable position. Yeah. And that's very difficult then, because if this technology like tricks them to believe the technology itself is like empathetic or whatever, or like in the chatbots, they're using avatars or something. So yeah, at least have something to project your emotions on and it kind of mimics them back. Um, it's ethically pretty difficult.

00:34:35

Caroline Bollen: Yeah. Yeah. Um, so, yeah, actually.

00:34:40

Caroline Bollen: We had a pretty interesting take on this, um that argued that instead of viewing that as manipulation, we should consider that fiction. Because when you're reading a book or watching a movie, you also kind of. When you are starting with that activity, you kind of accept that you are going to engage with something fictional. But for that period of time, like while you're watching the movie, you kind of accept that as the truth. So when you get introduced to all these characters, you. You know, you're not constantly thinking, Oh, this is a movie, These are actors. You kind of engage with the idea that this is true, even though it is not. And. And you can also engage with an emotional, an AI in that way. So for while I am using this, I am kind of accepting that this might be the truth. While it is fictional, just like when you engage with a book or a movie.

00:35:46

Caroline Bollen: Um.

00:35:47

Caroline Bollen: But that is quite an interesting take. And it also requires some different, a different perspective on it because, you know, we, we learn from a young age what it's like to watch a movie and read a book and that it is indeed fictional and how to engage with that. We don't, of course, learn that with these kind of things. But you could, you know, if you frame it that way and also if people are taught to engage with it in that way that that completely changes the picture. And we don't consider a a fictional story, a manipulation.

00:36:19

Me: Yeah that's right. That's really a very interesting take on it. I think it might be a bit problematic, like in like more gamified apps like Replica or something. It might work I guess because yeah you have like more control also to, to create your avatar kind of. Um, and then you have this more distant relationship or at least like even if it's not conscious, you know, it like, um, but like for those therapy apps or chatbots, I'm not sure because people are so vulnerable and I don't know if they would be reflective enough in that situation to accept this fictionality of it. And also like, um that there are real therapists behind it who develop like those responses. It's, it makes it a bit less fictional, I would say.

00:37:15

Caroline Bollen: Yeah.

00:37:15

Caroline Bollen: Um, then, and also in.

00:37:16

Caroline Bollen: These conversations you have your own experiences and the maybe the things that you share that are not fictional, so you get that strange blending. Um, yeah, but this is just like a different perspective because like from the perspective, you can indeed also critique why you cannot apply that like excuse to, to these kind of applications because you mix them reality and fiction in a very strange way, like within a conversation. Mhm.

00:37:49

Me: Yeah. That's really interesting. I probably have a section about like I want to include this topic of mystification and anthropomorphism of those apps. Um, probably it could fit in that. Yeah. Um, if you have an article or something you could send it to me.

00:38:09

Caroline Bollen: Oh, I'm not.

00:38:09

Caroline Bollen: Sure if it's published yet, so maybe it's not very useful.

00:38:14

Me: Okay, well, but.

00:38:17

Me: I can at least mention it.

00:38:19

Caroline Bollen: Yeah.

00:38:20

Me: Yeah, It's really an interesting perspective. Um, yeah, Something more you want to know.

00:38:30

Caroline Bollen: Um.

00:38:31

Caroline Bollen: No, I'm good. Is there something else you want to discuss?

00:38:37

Me: No, I think I asked everything. Yeah. Very interesting research you're doing, actually. Also very nice perspective you're taking from the virtual ethics side.

00:38:49

Caroline Bollen: Thanks.

00:38:51

Caroline Bollen: Yeah, I hope it was a bit helpful. I'm not exactly, precisely in the field that you're interested in, but I hope it still gave you some food for thoughts or anything.

00:39:02

Me: Yeah, definitely. I mean, I found no one who was, like, really doing exactly what I'm researching on.

00:39:08

Me: So.

00:39:10

Me: Everyone was just transferring knowledge. And I think that's that's fine. That already helped a lot.

00:39:14

Caroline Bollen: Yeah.

00:39:15

Caroline Bollen: Yeah. If you then still can find people like from different fields or different perspectives and. Yeah.

00:39:23

Me: Yeah.

00:39:24

Me: Thank you a lot for your time.

00:39:28

Caroline Bollen: You're welcome.

00:39:53

Caroline Bollen: Well, good luck with writing your thesis. Yeah, I'd be happy to read it if you finish.

00:39:57

Me: Yeah, I can send it to you, I think. I hope I will finish, like, end of August. Or at least I have my colloquium. End of September. I guess so. Somewhere in that time.

00:40:11

Caroline Bollen: Yeah.

00:40:33

Me: Yeah. Okay, great. Okay. Bye bye.

Interview Anna van Oosterzee

00:00:04

Me: Yes. Now it seems like it's working great. Okay. Yeah. Great. Um, yeah. As I said, I'm doing this research on, like, artificial emotional intelligence. Um, and the aim in this. Yeah. Track of research is to develop an AI that can recognise, interpret and process and simulate emotions. Um, and that ensures like more individualised and um. Yeah, more individualised approaches and interactions between the users and the AI. Um, would you maybe start by introducing yourself what you're doing in your research, um, and what links you to like AI and psychotherapy, psychiatry.

00:00:59

Anna van Oosterzee: So, Well, I'm Anna van Oosterzee. I'm doing my PhD at Utrecht University. Um, and I'm looking specifically at diagnostic AI, so AI that can help in the diagnosis process of mental disorders. Um. I've spent the first year mainly looking at the more philosophical literature around philosophy. See philosophy of psychiatry, because I'm going at it from a more ethical, philosophical angle and I'm now working on supervised machine learning and their implications on whether we can use this in the same way like oncology or radiology is using it. So let's put someone in a brain scanner, we put an eye on it, and then we ask whether they can see like, Hey, can you see in this brain if this person is depressed, yes or no? And it's a very non emotionally. So this is a nice contrast. Um yeah. So I've been I'm now in my half of my second year, so I also haven't been working on it this long. But my background is, as I just said in psychology, neuroscience, and I did my master's in philosophy and now this new development technology is just really exciting for these fields that have a lot of open questions, I think.

00:02:25

Me: Yeah that's right. Um, are you working on like specific applications or is it just like broader.

00:02:35

Anna van Oosterzee: So specifically these diagnostic supporting technologies. So I think. They are not in clinical practice right now. So it's difficult to give real examples because they just don't exist at the moment. But they're aimed more to support a psychiatrist in the hospital instead of being like a therapist. So supportive technology.

00:03:07

Me: So like in addition to usual diagnostics or in general?

00:03:13

Anna van Oosterzee: Yeah, I think it can be both ways. Like if you can develop something that's really better than the usual behavioural tests, like you have these really long questionnaires or patient has to go for observation for a longer time and this is just very unpleasant. So I think one of the aims is to replace that. I'm actually arguing in my next paper that I don't think that we'll be able to develop a technology that can do that. But I do know that the aim of the teams that are developing these technologies are also not not to replace like the interview that the patient has. So you're not going to talk with an AI, but you're still going to talk with the psychiatrist. And the psychiatrist has this tool.

00:04:01

Me: Yeah that's very interesting. Like you're working then on like a different angle of diagnostics, right? So like an additional perspective and both get combined, you said, Right.

00:04:14

Anna van Oosterzee: Yeah. You also have, you know more about than me, these apps and these applications where you talk with the AI and the AI does diagnostics. So I think it can also be integrated in these more all around intelligent AIs, but now they're mainly image classification systems.

00:04:39

Me: All right. Yeah. So it's not like speech based or anything?

00:04:42

Anna van Oosterzee: No, no, no.

00:04:44

Me: Are you like also looking for. They're like a lot of systems, like the FACS, which is like a facial recognition system for emotions. And it identifies like specific action units. It's based on the basic emotional approach of Ekman where you can identify like specific mimicking or whatever. And they said it can be also seen in the brain activity. Are you looking for similar things or is it just completely different patterns?

00:05:21

Anna van Oosterzee: Um, well, what I've been looking at is just the brain activity and I've been mainly criticising the DSM system and I think it's there's a lot of possibility with these technologies if you let go of the classification, depression or anxiety. I think that if we develop these wonderful like emotion recognising or mimicry systems, but then still try to diagnose depression that that's just not going to add a whole lot more system.

00:05:53

Me: Yeah, right. Yeah, I agree totally. Yeah. Like I also read a lot about the criticism about the DSM and those strict classification systems. And I also have the feeling like that the whole field is turning away from those strict classifications. Yeah, especially when you're.

00:06:10

Anna van Oosterzee: Frozen. Like you cannot move once you get into that corner, you cannot do anything with the classification depression. You're completely stuck once you adopt it.

00:06:20

Me: Yeah, right. And also a lot of people like get stuck in those stigmas and everything. So yeah, it's better to do it like more individualised what's at least the promise of those technologies. Also, even though it's not that clear if it will work out because it's still rule based, so it's hard to turn it away from those strict classifications, but they at least vision or promise it to do. Yeah, but.

00:06:46

Anna van Oosterzee: I can imagine that there you can do mirror interventions like on a small time scale. So not the whole diagnosis. Like you have anxiety or you have panic attacks. No, but then like the moment of the panic attack, like if that can be recognised or you can train with it like, Oh, now my anxiety is rising, how can I control this and can have the feedback of like lowering it again? Then you can stay away of from those diagnostics that are not really helping you, but still. There have a lot of benefits for the patient.

00:07:17

Me: Yeah, right. Like in the immediate moment where it's happening. Yeah, I agree. Um. Okay. Let's move on to the next question. Um. Yeah. Now that we talked a bit like about the possibilities that are probably there and what you're working on, um, like where do you see the most potential for these Technologies. It kind of catches up with what you just said, but yeah, what are the areas where you think are the most? Potentials and possibilities where could be implemented.

00:07:55

Anna van Oosterzee: I think on one side there is the immediate treatment possibilities. So just enhancing the knowledge that we do have right now been looking at a Start-Up that uses an algorithm to predict which kind of antidepressant will have the most benefit to the patient. And I think these kind of little tools, like these little tricks, can just help a whole lot, because for patients it will take like three years to test five different antidepressants. And if an algorithm tell you that in five seconds, like for science that's quite simple or maybe not the most exciting type of AI, but for patients that's really what you need. So I really like that. I like the prospect of collecting a lot of data. Now with the wearables and your iPhone, you can really collect these fine grained longitudinal data, which before was not possible. Like I remember when I started my study of psychology, there was all this complaining of, Oh, you can't collect the type of data we need. And now ten years later that's just not yeah that's not the case anymore. You can collect the type of data you need and you can process this big data sets. It's amazing how fast this is developing. I think at the end of my PhD, like the first things I said in the first year already, I'm not correct anymore. Um, I think that I'm personally very interested in is a more philosophically also exploratory is also. Um. Oh, no. Lost word. Um, these really large neural networks, they, um. Oh I'm writing in Dutch. So now my English is really stuck. They show that they can perform actions and have this additional level of features that we didn't really expect it to have.

00:10:15

Me: Like all the layers.

00:10:16

Anna van Oosterzee: Yeah. Um. Well, we'll get back to that. Let me think about it. So I will come back to that after the questions, if I. Okay. Word.

00:10:38

Me: Great. Yeah. Um, then probably you also addressed, like, ethical sides you're addressing, like, what are you focusing on there?

00:10:50

Anna van Oosterzee: Um, one of my worries is that with developing these technologies, using biological, using like using this AI, you can have. You can give patients a sense that you really know what you're talking about that you really found what is wrong with them. And you already mentioned the stigma. But I'm even more worried that the patient itself will believe like, Oh, see, ADHD really is a thing. And that's really wrong with my brain and this is why I'm the way I am. And. Well, we know that this is not the case. We know that ADHD, the way we use it right now, is not a thing. It's a explanatory model. It's not a causal structure. It's nothing of that. But you can't explain that to a patient, especially not when you have a brain scan and an AI that's, oh, you're really have ADHD. But. No, no, you don't. And I think, yeah that for me that's lying to a patient and I have issues with that.

00:12:01

Me: So it goes along like with wrong or probably incomplete diagnosis or labels put on the patients and they kind of internalise it.

00:12:12

Anna van Oosterzee: Yeah. And it's a false sense of security almost. Um, you, you try to sell it to a patient that might be doubtful, um, with all these tricks. And I think the psychiatrist themselves also think that it's more reliable than it really is. And of course, the psychiatrists are also not experts, so they don't really understand the tools, even if they could read a brain scan, for example. They are also not fully aware of what exactly they're implementing in their own process. Mm hmm.

00:12:49

Me: So you're worried about, like, overreliance on the tools and like that they provide kind of wrong security for the patients and the therapists? Yeah. Yeah. That they. Yeah. Think they have something to rely on, even though it's not like very true. And you cannot ever probably say if it's really true. Yeah. Mm hmm.

00:13:14

Anna van Oosterzee: Yeah. Yeah. And for example, a lot of people talk about biases or privacy and data security. I suspect that most of that will be resolved at some point, like it needs some work. But we know the work that needs to be done and biases are problematic and tough and structural, but we also are aware that they exist and that we have to deal with them. So in a way, I feel like those are easier problems than ones that are really integrated, integrated in the system.

00:13:54

Me: Do you see any other challenges that might arise with the use of AI?

00:14:01

Anna van Oosterzee: I'm generally quite positive. Like I like all the tools that are being developed, like all the extra possibilities that patients get. Also because it gets like it gets more dimensional. If you have problems talking to your to your therapist directly and you also get an app or a tool that's more fits your lifestyle, fits your structure, like all of that, I think that's fine.

00:14:36

Me: So you think like the whole approach to mental illnesses will get like more layered and nuanced by using different sources of information.

00:14:46

Anna van Oosterzee: I would hope so, yeah. Of course you always have, like the capitalistic movement that will push out the most expensive thing, which is the human labour. So you need a government and healthcare system that protects like the true therapy. Well that's that's being squished now anyway. Like we're having trouble deciding how much money needs to go there and everybody is yelling like, oh, there's not enough help and there are not enough places where you can learn to become a clinical therapist. So I don't think that's an AI problem. That's just a capitalistic health care problem.

00:15:38

Me: Yeah, right. I think that issue arises always. Um, but like also, um. There's actually a lot of research going on in this field. And you also said like there are basically no real applications currently, so it's not used in practice. So how would you explain that? Or do you have any clue why that is the case?

00:16:04

Anna van Oosterzee: Well, I think a lot of those. Smaller helper technologies. I'm not writing about that, so I wouldn't know how to call this. They are being applied to clinical practice. Um, my friends who have a Start-Up and they have these VR headsets that give EMDR therapy so you can do the EMDR therapy in like a 3D setting. And I think that's really cool and really innovative ways of improving therapy. For diagnostic I think the problem is that there's nothing to gain there. It's not moving to clinical practice because we have no idea about the biological mechanisms so we can train an AI to discover that for us. Um. But it's much more difficult than, for example, with cancer screening. They know what they're looking for. They know they already have doctors that are trying to find those specific patterns. And you're just automating this pattern finding process and you're already struggling, of course, to get high enough accuracy or liability to really make it to clinical practice. But they're moving there. And for psychiatry that's just not the case. Like, we don't have psychiatrists that are diagnosing on brain scans. We're not automating the brain scanning process or trying to create a whole new oh, no. Now maybe finally we can say what we're really looking for. And I believe that AI is just not going to or at least machine learning is not going to give the answer that we're looking for. So it's never going to move to clinical practice for this type of technology. I hope so. Because of, well, the ethical concerns of creating this false reliability.

00:18:07

Me: So you say basically like that they're developing technologies that are not even useful or they are useful but not accurate. I mean that's a very harsh statement.

00:18:21

Anna van Oosterzee: Yeah, I think they're chasing a dream and not realising that that's not the way to chase it. I think they're really hoping that I will resolve some really tough problems in psychiatry and that were slowly realising now that that we can't resolve it in this way. Like I hope we can resolve it, but not with this specific tool.

00:18:47

Me: Basically because we're lacking like the scientific evidence from like other fields, like biology, as you said.

00:18:54

Anna van Oosterzee: And maybe we will never find it. Like if you believe the non-reductionist, they will say, Well, of course you're not going to find this because mental disorders are not in the brain. So stop looking there. Like just do that. So as long as that question is still open. But then, for example that I met the Antidepressant Medicine AI tool, I do think this will move to clinical practice soon.

00:19:24

Me: Okay, So there's like two sides of those applications, like one that are very idealised and putting a lot of vision into them, even though there's like a spark of knowledge that it won't work out. Yeah. And then the others who are really have like, realistic possibilities. Yeah. Yeah. For getting along. Okay. Um, how would you, like, draw a line between them? Like, what types would fit into which category?

00:19:54

Anna van Oosterzee: Um, so I think the types that really focus on improving prediction. So that's the DSM Categorises categories are not really predictive. Um, and so when we focus on more reliably diagnosing, we're not really improving that predictive value. So I suspect that if you focus on improving a specific, specific predictive value, so for example, treatment effect, um that there we can make the way and um, well, for example so that um, you have AI that flags Facebook posts based on suicide predictions. So this is also specifically like preventative prediction, like, okay, can we, can we predict this patient might be a risk for themself? We don't care what kind of disorder there is. We just care that right now they're they are at risk and we have to catch them somehow and treat them preventatively before something happens. Probably they already have depression and such or such, but the AI is just protecting this one specific type of behaviour. And there I think there's much more promise and also much more interesting ethical dilemmas about if you predict it wrong or do you have the right to use all this data and intervene in your daily life? And they also get into discussions like, Okay. How much is social media a private space or public space control? Would a governmental health care program have there? Or how much responsibility do they have for their own users? So I think there things will start getting interesting.

00:21:57

Me: Yeah, right. You also mentioned like the problem with data. Again, in the beginning you said that you think one possibility is that you are able to gather all this data now, but like from regulations and everything and also the ethical directions, it's more like about privacy and that you cannot really access it. And that's what also mentioned in the literature often about like why those technologies and psychodiagnostics are not that far because the data is not there or cannot be used. Um, so that's a big issue in the development also. Um, do you think there are like developments in the direction that you can finally access this data or use it in?

00:22:45

Anna van Oosterzee: Well, it is annoying that privacy laws are blocking research. It's a good thing. Also, I think that privacy. For me, privacy is not the most important ethical concern. So sometimes it's frustrating to see that privacy gets put on the first position and then blocks other, more important ethical concerns. So if you ask me, I think preventing suicide is more important than privacy. So I think that an algorithm that could prevent this, like a real good one that's proven effective. Should have the right of way over a privacy law. And of course, there are these laws that if there are real important reasons to. Sorry, my thing will stop it. Um, you can block it. Um, well, you can also with the Fitbit and your iPhone, you can, of course, just give access. Um, and even though on the one hand, privacy is blocking this access. Ten years ago, there just wasn't even the technology and the data to be blocked. So we have to figure this out. But there are at least are opportunities now that weren't there before. Um, I don't know about if this is the reason why things are moving slowly.

00:24:20

Me: Probably not the only one.

00:24:23

Anna van Oosterzee: Yeah. Reasons also. Was reading about, um, a chat function that acts like a therapist. And there they were, debating on how much data you should add to it. So how much clinical data you should add to it. And surprisingly, not even that much because it's such a rich network that you don't even need to add that much data to really train it on specific emotions or specific, um, kind of therapy. So I think that which is really, really, really large models, extra data becomes less of a problem. It's like only. A first step development problem. You need you need the huge amounts of data to make the profiles. But once you have the profiles, you need a small amount of data again. Mm. So right?

00:25:30

Me: Is there actually like a real function within ChatGPT for this therapy thing? Because I just read that people are like misusing it as a therapist, as a personal therapist, kind of.

00:25:41

Anna van Oosterzee: You have all these add ons now. You have all these apps that are just fronts for the questions that are inserted into ChatGPT. I've never used one, but they're everywhere. Suddenly, like every week there like a thousand new versions and some of them do therapy I use, um. What is it called again? Very cute little app. Voidpet.

00:26:09

Me: I heard about that. Yeah.

00:26:11

Anna van Oosterzee: And they also have a chat function where you can chat with pets and just express your emotions. I don't think it recognises emotion, but it still responds in like a pleasant manner. Yeah, it's already used. It's used everywhere. But this one was the paper was reading was really about developing therapy technology. And less of this emotional support. Animal technology.

00:26:43

Me: Yeah, I can see. Um, yeah, you you mentioned emotion. Yeah. Yourself. Like, do you think it could add something to, like, the research going on if you're adding, like, those emotion recognition, emotion mimicking functions into the AI system?

00:27:05

Anna van Oosterzee: Well, those are two different things. Of course, you have the emotion recognition. Um. Which I think. I would mainly be interested to see how it would learn those emotions, how it's labelled, how you classify the different types of emotion. Also, because we as humans, which are knowledge about emotions, are struggling. Okay, what are really markers for which emotions? Um. Get again. Measuring distress or arousal is fairly straightforward. Mimicking emotion, I think, gets into like it's a bit murkier. Yeah, I think you have a lot of uncanny valley risk with that. And also don't know if it's necessarily necessary for a therapy tool like a therapist is also not. There to have a lot of empathy with you. The therapist is there to help you work through your problems, not quite themselves. So maybe an AI doesn't have to mimic that much emotions. If it wants to be an effective therapist, it just has to mirror what you are doing and to help your process through your options and your biases. But in a way, a therapist is also a bit empty. They're not really supposed to get their own emotions involved into the mix.

00:28:42

Me: They're not probably involving your own emotions as a therapist, but at least like some kind of reflecting on the emotions of your patient and also probably a bit of reassuring. And I think that at least requires a bit of empathy.

00:28:58

Anna van Oosterzee: Yes. Yes. So it's.

00:29:01

Me: Not just think like.

00:29:02

Anna van Oosterzee: Let's think like a real social interaction is in a way much more complex than a therapy session and requires much more emotional mirroring and interaction. While a therapist. Yes. Has to offer safety and respect and some empathy, but also only some empathy. Like if someone has really self-destructive tendencies, you can't be too empathetic about that. You can't be too distressed about that. As a therapist, you have to deal with the problem at hand. And so, yeah, I think don't think you need very advanced emotional mirroring for an AI. I think that would make a less effective therapy. AI if it really is going to say, oh, it feels so sorry for you and I feel your pain because it doesn't doesn't feel sorry for you, it doesn't feel your pain. It's there to help you process, but it's not feeling anything.

00:30:10

Me: Yeah. What about like if you're looking at it the other way around, like I read about AI tools that should be used in like to assist therapists developing better therapy skills like that they can interact with the AI to um, yeah. To get like.

00:30:34

Anna van Oosterzee: The AI is the patient, right?

00:30:36

Me: Uh, no, not really like that. But they can like if they have a client or patient who has like severe illness or something and they don't really know how to communicate it to the patient in like a sufficient manner, and then they can interact with the AI to train it or get like tips from the AI to like how to bring it to the patient and also how to deal with it themselves because it can be also like burdensome for yourself to have a patient with those diagnoses and not really knowing or at least I mean, there will ever be patients which you cannot really cure, like there will ever be like 1 in 100, I don't know. But at some point you will experience like patients you cannot really help in the end. And it might help like a therapist to deal with that. Also to not like take all those emotions to yeah, to home and also deal with it the whole day in your free time. Um, yeah.

00:31:41

Anna van Oosterzee: So. But wouldn't that be more like a place to vent for the therapist then? Because I still think that to respond very emotionally as a reply, like the reply doesn't have very emotionally it, it needs then I guess a lot of theory about therapy and some like support, but I think similar. It's just therapy, it's a method, but emotionally it's just therapy. It's there. Like, yes, it's a safe place. You need to get it out. But I'm not here to enter with you into this emotion. I'm here to help you guide through your own emotion. I'm not there with you in that emotion. Um. But I think that's I haven't read anything about this kind of emotionally standpoint for therapy. And therapy, I think would be interesting to to see if some more can be said about this. Well, slight lack of empathy in therapy, because I believe that that's a better type of therapy than a very empathetic therapy.

00:32:51

Me: Yeah, right. To like, keep the the professional amount of distance to your patients. Yeah. I mean, it's also written in all the ethical codex and stuff for the therapists.

00:33:01

Anna van Oosterzee: Or like, at least an appropriate amount.

00:33:04

Me: Yeah. Right.

00:33:06

Anna van Oosterzee: Of course you don't want an AI that's harsh or cold or uninterested, so you do have to give. Yeah, you do have to give your your vibe of sympathy. But just a vibe. Just a hint that's already there. Interesting. Interesting thing to see if you can create it and how much emotion mimicry you need for that.

00:33:30

Me: Yeah. Right.

00:33:31

Anna van Oosterzee: And also how much emotion recognition you need for that.

00:33:37

Me: Yeah. I mean, they're like a lot of models, like some work with, like, image recognition and facial expression recognition, but some also work via natural language processing, audio processing that they get, like those verbal cues of emotions. But still, as you said a while ago, like there's no common sense about what emotions are, how they are expressed. And that's also a huge point of debate, like the basic emotions by Ekman were criticised so severely.

00:34:08

Anna van Oosterzee: And it's the same like the categories, like the emotion categories also are problem here in this system.

00:34:17

Me: Yeah, right. It's like not graspable and it differs from person to person, from interaction to interaction between cultures. And it's so different that you cannot really yeah. Draw a clear cut line. So I have the feeling like they kind of instrumentalize this approach to make it work for AI systems.

00:34:38

Anna van Oosterzee: Yeah.

00:34:39

Me: But like, yeah.

00:34:40

Anna van Oosterzee: They try to simplify a very complex problem and I think a lot, of course it's interdisciplinary research and you have to simplify both fields a little bit to bring them together. But it seems like in both cases, sadness. Well, as a philosopher of emotion or psychologist that focuses on emotion, you know that sadness is not a simple thing. So you have to somehow explain the computer scientist, what they have to label sadness for supervised or recognising pattern recogniser to recognise. So I think there's some something lost in translation between these disciplines of the complexity of these kind of theories.

00:35:28

Me: Yeah, right. Definitely.

00:35:33

Anna van Oosterzee: Um.

00:35:35

Me: Yeah, probably. You can tell something about like, um, what your vision is for the field or what you're hoping for since it's a visionary field. Um, and we cannot really know what to come. Like, what would you want to come?

00:35:54

Anna van Oosterzee: Now?

00:35:57

Me: Or probably want not.

00:36:02

Anna van Oosterzee: Well. So one of my worries is that these. Simplified, complex theories get locked in so that as a society, we're really going to rely even more on these simplified versions of sadness or depression and that that's not really helping a dynamic development of our knowledge and also our societal opinion on stigma and depression and ADHD. Like it's moving so fast. And if you're going to implement systems that really rely on a specific definition of these kind of things, these kind of concepts, then you might also limit some development that that's also going now, like the whole neurotypical neurodivergent movement. Doesn't want an AI that's trained on a specific description of autism because it's moving very fast right now. Like don't put that in into a very expensive. I just yeah. Let it flow or at least. Leave the flexibility for that concept to change, because scientifically we don't know what it means. As society it's very important to know what it means. So there have. Yeah, like this log in problem and I just hope that it becomes more personalised. So all these smaller tools can help a lot with, uh, self-development, self-expression and helping you deal with your day to day emotions also because, well, a lot of research shows that that's just the most efficient thing that we can offer at the moment, like we also don't really have good therapy that really resolves deep issues. Unfortunately, we don't have that to offer. So if we can make the smaller therapy more accessible also to a lot of people that might not be able to go to a therapist or while the system is very complex, you have to go to your your general doctor. You have to communicate clearly what your problems are. There's so much gatekeeping in this whole process for someone to actually arrive at the doorstep of a psychologist and then explain to their psychologist what's going on. So I think if there are a lot of smaller mental health support technologies in world that that's just a good way to go at this moment. So yeah.

00:38:47

Me: Yeah, right. Um, so in general, what you say, like AI tools and psychiatry are desirable or rather not.

00:38:58

Anna van Oosterzee: I think the whole field of psychiatry is not very desirable at the moment. It's not the AI's fault and it's not the problem here. But, um. No, I think. Um, it should not get in the way of real human interaction, but. Real human interaction is not always desirable or not always possible.

00:39:28

Me: Okay, So you say they would supplement each other? Kind of.

00:39:32

Anna van Oosterzee: Yeah, I think just a lot of a lot of space for supplement. Yeah. For for cohabitation of technologies and humans. A lot more than some people are afraid of. I don't see them replacing psychiatrists psychologists. I do hope they replace psychiatrists because psychiatrists. Um. But like proper psychoanalysts or therapists, I think are still very useful and necessary, but in a more diverse way With like with these glasses, if you can talk to your therapist once a month and then at home in the same environment, try to work through your trauma with these glasses, for example that just you can have a lot more therapy than you could only have with going to a therapist once a month. Yeah, nice. Quite positive about that.

00:40:34

Me: Yeah. Great. Okay. Um, do you have, like, any comments or anything you want to share, which we haven't addressed yet?

00:40:45

Anna van Oosterzee: Still searching for the word. Let me. Let me. Um. Emergent properties. That's it. Okay. That's a word to lose, right? Yeah, Right. Now, so that's what I find most exciting about this development is that you can see now with ChatGPT is so large that you get these properties that are not really predicted and a little bit out of our control. And I think that mirrors a lot of our mind, which is also very well emergent properties is now the best reductionist explanation we have. But now we have a tool that does the same kind of thing and we haven't had that before. We haven't created anything that's also kind of emergent. So I think that's really exciting thing to start exploring this. And like this is this year, last year this happened, um that that these networks became so large that this really started happening. And so I think for the future, future that this is the place where real development is going to happen and real possibilities are that that can just take shapes that right now we can't even imagine that they don't think are going to be super human like, but might add more dimensions to our society and our knowledge base and also our understanding of herself that. Are now difficult to predict or see coming. So that's where I would like to focus my attention on, see what's going to happen in that very weirdly philosophical, technical era of AI.

00:42:42

Me: Yeah, it's very interesting and very complex as well.

00:42:46

Anna van Oosterzee: Yes. Yes. Try reading articles. I'm like, okay, Chaos and Emily. Emily Alienation. And so it didn't study mathematics. So then this is good. Yeah, it was tough.

00:43:02

Me: To understand sometimes.

00:43:04

Anna van Oosterzee: Yeah. But yeah, I think that's so the development of AI itself is exciting of these networks themselves are exciting and think the applications will come. Mm. Sort itself out.

00:43:20

Me: Yeah, right. Um. Yeah. Do you have any questions left? Probably.

00:43:27

Anna van Oosterzee: Uh.

00:43:28

Anna van Oosterzee: I'm kind of curious about your opinions. I think you heard a lot of different interviews. Had a lot of different interviews with a lot of different. Uh, yeah. How do you feel about this whole development and possibilities?

00:43:44

Me: Yeah, like when I started, I was, like, completely against it.

00:43:48

Anna van Oosterzee: Really?

00:43:49

Me: Yeah. I was feeling like you cannot implement, like, AI in psychiatry at all. It's like, insane. Those people are so vulnerable. And yeah, it's you simply cannot just put a system on it that's so unsensitive. But like, the more I read and the more I got presented to all those possibilities that are actually there and that actually also have proven to like be efficient, at least in the laboratory setting. Like, I mean, it's not proven that it works the same in the real world then, but they are pretty convincing. But also like that makes me sometimes forget about the pitfalls that might come with that because like it's like presented by companies most of the time. So yeah, they want to sell their products and they want to advertise them. So you really have to be aware of that and careful while reading everything and also reflect, Yeah, from your own perspective about it. But I'm not that negative anymore. I would say. I would say there are actually possibilities and there are a lot of reports about people like using those chatbots or something, um that really helped them. So I think that's nice, even though probably I wouldn't use it for myself, but if it helps them, like I also had a psychiatrist in an interview and um, he told me like he was not working with AI technologies but with VR technologies. Um, and he said that like the patients are really open to those technologies because they want to be helped and they don't care how they are helped, but that they are helped. So that also kind of changed my mind a bit. I mean, if they really want it and if it helps them, it's fine. It's great actually. If you have more possibilities then. But like.

00:45:40

Me: If it works, it works.

00:45:41

Anna van Oosterzee: Yeah. Right. And it wouldn't work for everyone because yeah, everyone is different. So yeah, I think if people are open to it and want it, it's totally fine to do it, at least if the systems are validated and reliable. Yeah. So yeah, and that's that's another issue like all this research going on, it's, it's tough and it's also like a bit stuck and like one researcher told what kind of you also told that there are systems developed which will just lie around and will not be used. So that made me worry a bit like it seems like such a waste of fundings because they get like funded projects, develop those technologies and in the end, no one really cares about it. Like I would feel that it's kind of a waste of my lifetime or my sources, and you could have used it to develop. Yeah, more helpful technologies or do whatever.

00:46:38

Anna van Oosterzee: It's not really linear in that way. Like you have to try trial and error. You have to try what works and also what society is ready to adopt in a way, you can't always guess that upfront when you're trying, you know, if just no one is like everybody says, like, Oh hell no, I'm not going to use it. Well, then it might just work just fine. But you're not going to sell it. Yeah.

00:47:05

Me: Yeah, right. Yeah. But he said, like, kind of. There's also a lack of responsibility, I would say, from the people developing those tools because they simply don't care about like anyway they could probably make it acceptable to society or at least make attempts to implement it kind of. So they just leave it if the funding runs out and then it's there, but no one cares about it anymore. And that's also like kind of missed possibilities, I would say, because if they have been proven to be effective but never like have been really tested or really any effort has been put into. Yeah. Implementing them then. It's actually a pity. Yeah. But yeah.

00:47:50

Anna van Oosterzee: And the generalisability now is also a huge problem I think all the time that then maybe it works for like a small subset of people, but then trying to get it to work for a large set of people. These types of networks don't like that. Hmm. And it's funny. It's weird. It's very human in a way. It it's stuck in a pattern. It doesn't like to change anymore.

00:48:22

Me: Yeah, right.

00:48:24

Anna van Oosterzee: And AI is a bit so did induct you say wat de boer niet weet, eet hij niet. So it means like, what if the farmer doesn't know a food, it won't eat it? Oh, yeah.

00:48:39

Me: We have the same thing in Germany.

00:48:41

Anna van Oosterzee: Because AI can be a bit like that. Like, I don't recognise this pattern. Like this. Yeah. Interesting. This is such a problem right now. It didn't see that coming. No one saw that coming.

00:48:59

Me: Yeah, right. Yeah.

00:49:02

Anna van Oosterzee: Probably they will have resolved this in two years so. Yeah.

00:49:07

Me: Might be. Okay, great. I think that was it from my side. If you have nothing to add.

00:49:17

Anna van Oosterzee: No, no. Just really cool that you're working on this. And I'm curious if you finish your dissertation. Would love to read it. Yeah, if you want. If you want to send it, Yeah.

00:49:31

Me: Sure, I can send it to you. I think it will also be like openly available at the UT website. Whatever.

00:49:38

Anna van Oosterzee: Oh, you can publish it.

00:49:41

Anna van Oosterzee: No, but they like upload all the theses you write there. So sorry. Yeah, but I can send it to you. Definitely. I hope I will finish it like end of August or September.

00:49:54

Me: Oh, you're almost finished.

00:49:56

Anna van Oosterzee: Yeah.

00:49:58

Me: Well.

00:49:58

Anna van Oosterzee: Good luck with the last bit and your job hunt. Yeah.

00:50:03

Me: Thank you. Yeah. Um, do you know anyone who might also be a possible interview partner? Like working in a similar field?

00:50:21

Anna van Oosterzee: Carolina. But you know me through her, so. Yeah. Um. That's how much?

00:50:39

Me: Okay. Yeah, no worries.

00:50:42

Anna van Oosterzee: No, no, I'm. Thinking maybe that the guys that are developing the EMDR tools try contacting those. I haven't spoken to them for a while, but they're nice. You could try reaching out to them, but I don't know if they're going to reply quickly enough and have a lot of time. They're called Psylaris and Richard. I don't even know if he works there anymore. You know, like with these start ups, like they fire everyone all the time. So I have to check if. In. Oh, yeah. No, no, no. He still works there. Oh. Said CTO Christophe Green-white. I'll put it in the chat.

00:52:18

Me: Great. Thank you.

00:52:20

Anna van Oosterzee: So you can just contact them through the website. I'm not sure if they're going to reply because he's always really busy. Also never replies to me.

00:52:31

Me: Yeah, I had the same. Like I tried to contact some companies who are working on such tools, but I never heard back of any of them. No.

00:52:39

Anna van Oosterzee: So you could try them. And what they do is really cool. And they started small and they also studied psychology and they really had this aim like, oh, cool technology and let's help some people to try them. But no, I think you're you came from the corner of people that I know. So yeah add so yeah.

00:53:02

Me: How many. Thank you.

00:53:03

Anna van Oosterzee: Anyway, how many people do you still need?

00:53:06

Me: I think I just need one interview like I wanted to have eight, but I reduced it to five because it's so hard to find interviewees like no one is replying, no one is showing up and actually no one is really working in this field because like initially I wanted to interview like psychiatrists or psychotherapists really using those apps, but there are none at all. So yeah, I had to do it like a bit more speculative and also include more researchers. Um, so yeah, it's very tough to even find or figure out people who are working in this area who really reply and are willing to be interviewed.

00:53:43

Anna van Oosterzee: So no, I also noticed that when I started with my PhD, I was like, okay, let's first read all the literature there were like five papers at that point.

00:53:53

Me: Yeah, it's insane. Like everyone's talking about it and there's so much literature out there, but like you cannot really figure out the people, even though, yeah, you would think there's a lot, but there's not.

00:54:07

Anna van Oosterzee: It just isn't. Um. Yeah, I have a therapist friend. I know she, she uses some tools in her work, but also not really AI they're just not using it yet.

00:54:24

Me: Um, yeah, but that would also be an option. Like with this psychiatrist, I also did, like, in a more speculative way. I mean, if she's open to, um. Yeah, at least think about how it could be implemented. It would also be fine.

00:54:39

Anna van Oosterzee: Um. Yeah, you can. So Ruth house. She likes to talk about at least. So it's great. She's just a psychologist, so she's not really into the technology or the AI, but she's really interested.

00:55:02

Me: And yeah, at least then she would have heard about it. Yeah. And have an opinion about it. Yeah, I think that could also help.

00:55:10

Anna van Oosterzee: Yeah. I was teaching a course on technology in health care and then she was also one of the professionals. Oh, nice. So she does like to do these kind of great but academic tours. Yeah.

00:55:27

Me: Well, at least try to contact her.

00:55:28

Anna van Oosterzee: Yeah, I think there's a higher chance that you actually manage to speak to her within a week then with Kristoff. Yeah. So. Oh, great. Well, hopefully you get your last interview and then you can graduate. Yeah.

00:55:45

Me: Thank you. And thanks for your help and willingness to share your experience.

00:55:49

Anna van Oosterzee: Of course. See you.

00:55:51

Me: See you. Bye.

Interview Marloes Veldhuis

00:00:05

Marloes Veldhuis: Yes. Okay. Um, would you maybe start by introducing yourself and like, what links you to the field of psychiatry, psychotherapy, and probably also what could link you to AI.

00:00:22

Marloes Veldhuis: Uh, so my name is Marloes Veldhuis. I have been working as a psychologist for about five years now. First, at one practice, and now for almost three years at another practice, uh, called Mens GGZ, set in the south of the Netherlands. And I work as a master psychologist, so I don't have my BIG registration yet. I'm not a KZ psychologist or clinical psychologist. Um. I work with all kinds of different patients. I work very broad, so I'm not specialised in one kind of area of psychology yet and with different treatments. Um, I don't have much experience with AI in the field, but I do have some experience with e-health and I don't know if that's in any way comparable, but yeah.

00:01:19

Me: Mhm.

00:01:20

Me: Yeah that's very interesting. Like I had some interviews. You also, um, more related to those. Um. Yeah. E-health applications because AI is not really used yet, so that's totally fine.

I think you can transfer some knowledge at least. Mhm. Um. You're right. Anna also told me that you were like an expert or something. In a course. She teached. Is it right? Um.

00:01:50

Marloes Veldhuis: Yeah. The students interviewed us for, um, related to diagnoses and the ethical implications of diagnoses in the field.

00:02:03

Me: Yeah that's very interesting. Um, did you also talk about technologies there, or was it rather.

00:02:11

Me: General.

00:02:12

Marloes Veldhuis: No, not that much. I've also had some experience with experience sampling. I don't know if you are aware of that. In the past, in my master thesis I wrote a paper on um, using experience sampling to track mindfulness exercises and how it influences positive effects and positive cognitions.

00:02:38

Me: Yeah, right. Like what I experienced is that the applications who are already in use are mostly regarded to those more like wellness lifestyle apps and not. Yeah. Like specifically targeted to this psychiatric context yet because it's simply not allowed or or. Yeah right. And also like in general the technologies you can find are more like not so therapy related but rather like self-help apps or. Yeah, like wellbeing mindfulness apps. Meditation apps. Right. Um, yeah. But we could start to like include some kind of AI, do you know, like any specific applications or do you have like any imagination where it might um, be implemented in, in this psychotherapy domain.

00:03:40

Marloes Veldhuis: I have not yet have I don't know any particular applications. I have heard in the past about one of my managers once was speaking about the implication of using AI don't know if it was AI or a psychiatrist at a distance to quickly diagnose people, um, but don't know the findings of that.

00:04:07

Me: Um.

00:04:10

Me: What was it like.

00:04:11

Me: Quickly diagnosing like doing it faster than a human or just the the remote aspect or what was the point of it?

00:04:20

Marloes Veldhuis: I think it was mostly the remote aspect, and I think it also had to do with the the shortage of I don't know if there's an English word for this, but you know what it is in. So it's like in, um, in psychology trajectories, you always have to have somebody that enters possible over the trajectory and we'll call that. So I don't know how to translate it, but often it's either psychiatrist, clinical psychologist or psychology that will be responsible over all the. Yeah. The, the, the counsellors or the psychologists that are connected to one patient. Um, and there's a shortage of those people. There's big registrations. And I think that application had also been the goal of, um. Yeah, yeah. Filling in that gap and making it more easy to diagnose so that. Yeah. That's, is taken care of.

00:05:25

Me: Mhm.

00:05:26

Me: Yeah. Right. That's actually. Yeah. Catches up with what I found. Um, they are like um, promises that there will be applications who will help with this shortage of professionals. Um. And how to make it more accessible for people also like in remote areas where they yeah. Would have huge transport costs or whatever that prevent them from attending therapy or people who are simply afraid to to go in public or talk to like a human therapist that they might be more open to talk to. I because it's yeah. Perceived to be like not so judgemental. And that is more neutral because it doesn't really have an opinion itself. Um, would you agree with that or do you think it might be different?

00:06:22

Marloes Veldhuis: I can imagine. I think a lot of people are, for numerous reasons, scared of taking the step towards a psychologist. I see this a lot. Um, and I can for my own perspective, I can also imagine this. Um, and I think in that, I think any application or I would in that sense be more an easier step towards, um, psychiatry psychology, which I also think is a cultural thing, maybe in other cultures where technology is even bigger, for instance, Japan or something, I can imagine it might even have more of an effect.

00:07:03

Me: Mhm. Yeah that's right. That's also addressed a lot like this cultural aspect. Um, also that AI might be like more neutral culture wise and.

00:07:14

Me: And will be.

00:07:15

Me: More accepted kind of. But there's also like in return this challenge because especially like in the field of psychology, it's very difficult because you have to deal with emotions. And that's what I'm focusing on, like, um, artificial emotional intelligence and where you can like, or how this technologies can identify emotions and then deal with them or use them in a psychiatry context, because I think emotions are highly involved in, um, in this interaction of a patient and a psychologist. And also empathy plays a great role on this. Um, yeah. What, what would you think Like would this implementation of emotions, emotion recognition and mimicking emotions, um, how would that probably change the practice or make it like more, I don't know, acceptable for patients?

00:08:20

Marloes Veldhuis: Hmm.

00:08:21

Marloes Veldhuis: I can imagine it would have big implications because I think that's one of the pitfalls of e-health right now that still you you feel as as as a psychologist that you still have to keep an eye on it to see how it's going and to see if it's if they're responding. Well. And I can imagine AI might even do do more for that and might be able to recognise also patterns when patients are doing less well and yeah, might might be able to, to, to grab on to that. Um, but I think a very important factor is that in the field of psychologist psychology, um, we have to deal with quite big responsibilities because sometimes luckily not, not not often. You're, you're, um, you have to. Yeah, it can be life or death situations when suicidal thoughts or maybe psychotic things come into play. And I think that would be the yeah. One of the things where AI is very tricky because then who will be responsible if something goes wrong? Because now we have to document very well. For instance, if somebody is suicidal, we have to really document very well in our in our papers how everything that we ask and how they responded.

And if we checked everything and did everything we could do. And yeah, I can imagine that with AI that would be harder in a way.

00:09:52

Me: Mm.

00:09:53

Me: Like harder why or like when the AI is doing this.

00:09:59

Marloes Veldhuis: I mean, more harder in the sense that then who will be responsible if something goes wrong?

00:10:04

Me: Mhm. Yeah. Right. Um, yeah. Especially what you mentioned with all this note taking and everything. It's very time consuming. Right.

00:10:14

Marloes Veldhuis: Mhm.

00:10:14

Me: Um, and that's, yeah. That's why they also are actually working on like technologies or technologies who could, um. Yeah. Ease this process a bit and, um. Yeah. Take the notes for you and also, like, review the notes for patterns. So yeah, kind of what you mentioned that it might help in this area so that you have more time to actually do practice and probably also to work with more clients instead of doing all this. Yeah. Paper works.

00:10:49

Marloes Veldhuis: Mhm.

00:10:51

Marloes Veldhuis: And I think that also is is difficult in psychology that you have so many gradations of how severe a case can be. We work with not not as severe cases. So often people that suffer from minor to mild anxiety problems, depression problems, sometimes traumatic events. Those are also people that have still a lot of resilience and can also work by themselves a lot. But I think you have also cases who are way more severe, who are taking into care. Yeah, they have to sleep somewhere and maybe are suicidal or psychotic. I can't imagine that it might be then harder also for AI to to deal with that because you. Yeah, you really have to be on top of the patients.

00:11:42

Me: So you think those applications are more like would work for clients with milder conditions than clients with like really severe conditions?

00:11:54

Marloes Veldhuis: But I think it's hard.

00:11:55

Marloes Veldhuis: I think especially in the beginning, it's I think it's more safe to to to practice it with with milder complaints. But I think also, for instance, in. I think when when somebody is in a crisis, it's always good to have to have people there to to check. But I think, for instance, with people with bipolar disorders who will be suffering from mood swings for for a long time often, and you have to be medicated. I think they already make a lot of use of e-health and other things to to monitor their moods and to see when they have to to to change their medications. And I think then AI could be very helpful.

00:12:39

Me: Yeah that's right. Especially this monitoring aspect. Like there's this expression, I don't know if you know it, it's like the quantified self, um, where like people are tracking all their data and especially with the technological developments, it's get, it gets like way broader and you get way richer data about everything. Like you can track heart rates. But um, yeah, all your bodily functions which could be like indicators for some specific also Yeah, mental conditions. And then in addition now you can have like more opportunities with all the technologies developed to keep track of also like different aspects, also mental aspects. Um. So I think, yeah, it's hard because you also get a lot of data then which has to be processed somehow. Um, but you can also get a lot out of that.

00:13:37

Marloes Veldhuis: Mm.

00:13:38

Marloes Veldhuis: Yeah, I can imagine.

00:13:41

Me: Um, yeah. You talked about those e-health apps. Are you using any specific ones or.

00:13:50

Marloes Veldhuis: Uh, and right now we're using NiceDay. It's quite new. So uses it a lot for also fully online treatments in the past. I also use Minddistrict and Life. I haven't been satisfied with all of them, especially now. I think NiceDay is okay, but it's very CBT oriented and we also work a lot with acceptance and commitment therapy. So for me it's a bit of a shame that it's not it's there's not much acceptance and commitment therapy material in there, but I think it's good. And they also make a lot of use of tracking. So a lot of mood sleep, all kinds of of trackers and use that also to yeah. To guide the process more so that the therapist can can take a step back and they check every morning how the patients are doing. And um, I think, I think in that way it's nice but we see also see that a lot of patients don't take the time and effort and I think that's a shame.

00:14:55

Me: Yeah, right. But I actually contacted the company. I wanted to do an interview with them, but they did not respond. I think they are actually using or planning to implement AI at least.

00:15:07

Marloes Veldhuis: Okay.

00:15:07

Me: Yeah. Those are often like just minor automated processes, but it's still AI.

00:15:14

Marloes Veldhuis: Yeah.

00:15:15

Marloes Veldhuis: It's a shame that they didn't reply.

00:15:18

Me: Yeah, well, it's hard to get like those companies involved or get in contact because, like, no one feels responsible.

00:15:27

Marloes Veldhuis: Yeah. Yeah, I can imagine.

00:15:29

Me: Yeah.

00:15:31

Me: Um. Right. And you mentioned there are like shortcomings with those technologies. Currently, what do you think are like the the biggest gaps currently?

00:15:46

Marloes Veldhuis: Yeah, I think the what they offer, I think they're very different. For instance, Minddistrict is very program based, so they offer an anxiety program, they offer an acceptance and commitment program. So it's very structured and you can let it go a bit as a as a psychologist. NiceDay is more hands on. And I think all of them have different shortcomings in a sense that Minddistrict you it's not very flexible nice day is limited in the in the kinds of um yeah. Things that you can offer so it's more you can do maybe a thought of a behavioural experiments but I think that's that's about it. So Minddistrict in that sense has more to offer. Um, I think one of the biggest shortcomings is that patients don't reply. They really have to be educated about taking it into their own own hands. And I noticed, especially when they still have contact with psychologists, they think, oh, the e-health. Um, so I've noticed very often that they don't do much with it. And I think that's a shame.

00:16:57

Me: Yeah, I can imagine. Like because it's like the, the self effort you have to put in it. And as long as you have guidance along, it's probably hard. Um. Yeah. So you think or your experience is that in general the acceptance or the willingness to use those apps is rather low, right?

00:17:20

Marloes Veldhuis: Yeah. Especially when Yeah, when as I said, we are still contact with, with psychology and think for instance with PsyQ, who do only online treatments I think. Yeah, they use it so much that that patients are also expecting to use it and they are more educated on it. And with us it's more blended. So I think that immediately makes it more easy for patients to to not put much effort into it.

00:17:45

Me: Mhm.

00:17:46

Me: So it's like the overreliance then still on the psychologist and. Which prevents them kind of using them. The the apps. Yeah.

00:17:56

Marloes Veldhuis: And I think I might think that AI can be a solution also in that when when patients feel the warm connection through, I don't know if it will be possible, but it might be. And I think that's what we say now might miss in it. Yeah. That you don't really feel. Yeah the connection I think is a big part of psychology that works.

00:18:24

Me: Mhm. Yeah. Right. That's what they're working on especially in the field with the emotions where I'm looking on, um that they can establish this like real patient relationships. This rapport it's called I think, um. So that the patients feel very comfortable talking and feel understood. And they also found out that it doesn't matter in the end. Like if they have the impression there's like a human behind it who's looking or reviewing their data, then then they feel comfortable, even though it's just an AI who's responding, but in a way that looks so human, then they're totally fine with it. Um, yeah, but um, although you said the patients are kind of not really using those apps, um.

00:19:15

Me: If they do.

00:19:16

Me: Was it efficient or do you think you could better go without them?

00:19:23

Marloes Veldhuis: It really depends on the application and it really depends on the patient. Sometimes patients are really eager and they take it hands on and sometimes it just just doesn't fit them. And I think, for instance, I was a very big fan of Minddistrict. It's yeah, it's a shame that it's too expensive and we don't use it anymore. But yeah, as I said, I'm not a really big fan of, of NiceDay. And I think the definition of ehealth is very is very broad because if you just do a video call, it already falls under ehealth and Yeah. Mhm.

00:20:00

Me: Yeah, I can imagine. So you mentioned like costs are also a big factor here for like deciding which apps to use.

00:20:09

Marloes Veldhuis: Mhm.

00:20:11

Marloes Veldhuis: Mhm. Yeah.

00:20:11

Marloes Veldhuis: So not for me personally but for, for my manager. Yeah.

00:20:15

Me: Okay.

00:20:16

Me: So those apps are not like "sponsored" by the insurance companies of the patients or.

00:20:24

Marloes Veldhuis: No we just get paid by, by, by, by direct treatment minutes and we direct. So any contact you have with a patient. So if I would call or chat with a patient for only five minutes, I can I can charge five minutes, but all other things are not. We don't get. It's all in those five minutes. Maybe. I'm not saying so. The money you get. For instance, if I see a patient for 45 minutes, I get money for those 45 minutes. And that also includes then the administration time. So e-health is also I guess it falls under direct minutes when you have direct contact with the patient. So if I email or chat with somebody, I can. Yeah. Get money from it. And I would wonder how that works with AI. So but to answer your question, I was straying away. The platforms are not funded by the health insurances. They only fund the direct when we see patients. And we have to as a company, they are supportive of eHealth. So in your contracts with the health insurance, they can give you more money if you incorporate e-health, but you still have to pick the provider yourself and pay for that yourself.

00:21:46

Me: Okay.

00:21:46

Me: That's interesting because the the Apps I found are like mainly chatbots. Um. And I think they're partly, um. You probably have to pay for it, but then the patient has to pay themselves, I guess, because it's like supplement of a therapy. So you don't really have a therapist by your side. Um, but they're not publicly available yet, I guess. I think, or at least not in Europe. I think there was one app. I don't know if it was the Wysa or the Woebot, um, who's at least available in the US. And you're talking to a kind of avatar there and it has really good quotes on, um, how the patients reacted to it and felt understood. Um, and also in the treatment of like, um, PTSD. But I don't know if it were like lab trials or if it was in the real world. It wasn't really clear in the report.

00:22:52

Me: Um, but.

00:22:52

Me: I think that the, the patient has to pay themselves because they mentioned like or there's a free option and a paid option because they said like if, um, the like premium subscribers. Enter like a huge or frequent, overly frequent time of approaching this chat. Then they will be referred to a real therapist because it's like seems to be a more severe condition and they should get like. More professional help or like human professional help.

00:23:30

Marloes Veldhuis: Mhm.

00:23:31

Me: Yeah. So also it's, it's not clear like sometimes those applications work like as a, um, as a supplement for the therapist. So you work like with therapy sessions and do this in addition or it like take some tasks away like this note taking tasks they said at the beginning. Um, so there's always this connection with the professional, the human professional and the app. Um, but sometimes there are also like working independently or that's what they are working on now that it's supposed to be more independent and that they can be automated.

00:24:10

Me: Um.

00:24:11

Me: Yeah. What do you think? What would be like, from your experience more efficient?

00:24:18

Marloes Veldhuis: Uh.

00:24:19

Marloes Veldhuis: I don't know if it's, uh, would be less efficient. If it works, as they say, it should work. That's the AI is able to, to mimic and to support the patients. I think, um, I think it's mostly about the responsibility. And if psychologists or psychiatrists are able to, to let go in a sense, and to let those patients just go through the AI and not be, um. Yeah. Connected in a way. And again, I think it also really depends on the severity and about the what kind of problems there are. For instance in psychotherapy the connection is most important, but maybe in with EMDR, for instance, where it's, yeah, it's very mechanical already. It might be very easy for an AI to pick up.

00:25:15

Me: Mm.

00:25:16

Me: Yeah, I actually heard they're working on like, I think VR options for this EMDR. Um, so, and it seems to work pretty well. I think Anna told me about a Dutch company who's working on that.

00:25:33

Marloes Veldhuis: Oh nice.

00:25:36

Me: So that should work. And also like you could implement AI there in terms of like using avatars to interact with.

00:25:46

Marloes Veldhuis: Mm.

00:25:47

Me: So that's also a possibility. Um. Probably a bit more general like. Where would you see the the most possibilities for those applications? Like in which domains or in which specific application contexts?

00:26:10

Marloes Veldhuis: I think, as I said, in the milder, milder or moderate severity of complaints, but maybe also after treatment. So, for instance, in somebody who has bipolar disorder and was already been the medication has already been said that the after treatment and the follow up can be also done. I think those are the most important fields. And this the severe and the crisis. I'm yeah, maybe as a simplification, but I don't think primarily.

00:26:46

Me: Yeah, I agree.

00:26:49

Me: Um, and where would you see then like, probably challenges of applying those apps in the context of psychotherapy?

00:27:00

Marloes Veldhuis: You mean psychotherapy as in general or psychotherapy as in the really long treatments?

00:27:12

Me: More in general, I would say like in which domains it could be challenging or what aspects of applying it to like this psychiatric psychotherapy context could be challenging.

00:27:26

Marloes Veldhuis: I think being able to follow up the patients because when a general practitioner refers somebody, they also expect some kind of involvement. And I think for AI it might be also easier to track. Lose track of a patient. But maybe you can also think of ways to to deal with that. And I think also in the yeah, as I said, in the virtual responsibility, I think that's also a pitfall. But I think there are also a lot of good applications in Yeah. Making it easier for people to take the step, uh, for people with already a bit of resilience to, to take it into their own hands and to, to follow a treatment plan.

00:28:13

Me: Um, so you think it would be more advisable like for. Probably not using those applications or AI apps as a first contact, but then, um, rather as a psychologist assessing like the situation with the patient and seeing how open they are or how resilient they are. Um, and then decide if you want to use those apps or not.

00:28:39

Marloes Veldhuis: Yeah, I can imagine if there's a if there's a professional involved and if there is any responsibility. But I can imagine if there's not even a general practitioner involved or patients, people just want to, to connect with somebody. I can also imagine that they do it completely from themselves. But then I think the responsibility lies with with them. So.

00:29:03

Me: So with the with the patients themselves.

00:29:06

Marloes Veldhuis: Yeah. Only the company maybe that has. I think it's hard. Yeah. I think for the company that that makes the application, it's very hard. What are you going to do when somebody says, I'm really not doing well and I'm standing I'm maybe being dramatic right now, but I'm standing on top of a bridge and I'm. Who are you going to contact? Because you might not have any information at all and you might not even have the contact information of their general practitioner or. I think that in that sense, it might be dangerous.

00:29:38

Me: Yeah that's right. And that's still like an ethical issue that isn't resolved yet. I think that's also part or like a factor why AI applications are not yet implemented into this context.

00:29:53

Me: Um.

00:29:56

Me: You also mentioned like several times that this engagement of the patients is is also an issue. Do you think it will increase probably with the implementation of AI or with implementing even emotions in AI? So it's like more of a connection and probably also more fun because it's like more gamified a bit or because it's such a fancy technology. I don't know.

00:30:24

Marloes Veldhuis: Yeah, can can really imagine that if they feel the connection and if they feel. Yeah. Yeah. I think I think that's also a big factor why patients come to come to us because they feel connected and they feel also like when they make make an a, when they plan a session, they feel the need to to really come and to also do that for you in a sense. Um, so I can imagine that if you feel that emotional connection and it is more fun because I think they sometimes are just really boring right now. I think also with Nice Day for instance, yeah, you have to do this mood trackers but what. Yeah. What are you gonna. Yeah. It doesn't give any kind of reward to the patient in the sense.

00:31:08

Me: Yeah that's right. Like two other interviewees told me about this app. It's called VoidPet. That's pretty nice because it also does a bit of those gratefulness exercises and positivity exercises. Yeah, Yeah. And it's, it's actually pretty nice because they explain it's a bit like Pokemon because you have those pets and you can kind of feed them or grow them with doing those exercises. So there's more of this gamification and I think that's a bit more of what you said that would make it more interesting or is at least a more interesting design. And you probably have like also more motivation to do it then because you get like at least a small reward. Reward because those pets are like cute and you want to take care of them or whatever.

00:31:59

Marloes Veldhuis: Mhm. Yeah.

00:32:01

Marloes Veldhuis: Yeah, I can really imagine. I think I heard. Was it a patient? I thought they also talked about that and that they really liked it. And I can really imagine that. That it would help. I can also my boyfriend just walked in and he said to me about the movie "Her" you probably might have watched it. That I can also imagine that's an emotional connection with AI can also have very other kind of implications. Yeah. That you become too connected and it's also hard to let go and to to step into the real world again in some kind of sense. Because when would you say, I'm going to stop? Yeah, yeah.

00:32:42

Me: Yeah, right. They actually experienced it like a company from Berlin. And I think in collaboration with the UK, they developed a chatbot, it's voice-based, so it's like, like you call your therapist and then they. Yeah, just talk to you. And they're, they actually experienced this overreliance on, on the technology because it's available like 24/7 and you can always approach it and tell them what's going on and they just approach this app for like everything and every single problem they experienced for every single decision. And it's very hard to get rid of it then because it's also like it's available. For a therapist you have like a limited. Yeah. Number of sessions and then you have to disconnect kind of. But yeah, with those apps, it's pretty hard to to get rid of them. So there's like a bit of this addiction or overreliance problem.

00:33:47

Marloes Veldhuis: Yeah, I was going to make the decision then to say it's it's done now. Yeah. Because I already feel that that it's very hard with patients to set a boundary sometimes because they can get really attached and you have to be really, I think it's also healthy, these boundaries and I can imagine for AI that would be much more hard because that also might be sometimes, uh, money involved. And for the people making the app, it's nice that more people are and more and more are using it. And then where, where is the, where does it end.

00:34:21

Me: Mhm. Yeah.

00:34:22

Me: It's also like this issue with this capitalisation of this domain a bit because like people who are developing it and the apps who actually enter the market are more like company produced and with less of this research focussed back-up and they sell it as like wellbeing apps and not like in this psychiatric context. So it's easier for them to to enter the market and then they're already out there. So it's easier also to to probably turn them into those psychiatric apps. But still, then there are so many out there who are really lacking like scientific evidence and they are publicly and publicly available. So that's really an issue when people are just using it without knowing that there might be something wrong with that.

00:35:10

Marloes Veldhuis: Yeah, I can imagine.

00:35:11

Marloes Veldhuis: Yeah, it's kind of tricky. And then who's responsible in the end? Because I can imagine that those companies all have all kinds of disclaimer, uh, forbidding them from any, any responsibility in a sense. Mhm.

00:35:25

Me: Yeah, Right. Um. Also like in the end, it's kind of a more visionary field. So they're not really like real applications out there. Um, what would be your vision for this field? Like what applications would you find desirable and which not?

00:35:50

Marloes Veldhuis: I think from a perspective as a therapist, I would really like it to be my right hand and to be to be a support of what I'm doing with the patient. So that's kind of an extension of me and that I have some kind of control over it. But it also feels like it relieves me in a sense of of some burdens. As you said, it is note taking or following up the patient. Yeah. As a therapist that I really like that. And from a possible patient's perspective. I think it would be nice to have something that's easy, attainable. Um. I kind of can imagine that that you would like that. And if I wouldn't have any experience as a therapist, I might even, uh, imagine that AI would very be very easy step. But I wouldn't know about the how scientific it would be or, uh, yeah, where the boundaries are or Yeah.

00:36:59

Marloes Veldhuis: I can imagine.

00:37:00

Marloes Veldhuis: It would be an easy step for creation to use such an app.

00:37:07

Me: Easy. Because of what?

00:37:09

Marloes Veldhuis: Yeah, because it's.

00:37:10

Marloes Veldhuis: It doesn't as much of... You were so much on our phones already and I think for some people they associate their general practitioner right now with only medical things. And I think it's very hard to take the step to also discuss more psychological things with them. They have to take on the stoners also. Now that makes it a bit more easy, but still I see it all around me that this step is just very hard. And I can imagine that doing a three year phone feels much safer.

00:37:41

Me: Yeah, right.

00:37:42

Me: So you think patients would actually be willing then to use those technologies.

00:37:47

Marloes Veldhuis: Mhm. I. I think.

00:37:51

Marloes Veldhuis: Yeah, I think they would. And I think. Really Because they don't know. Yeah. Also don't know that are not aware of the risks or are not aware of what that might mean.

00:38:01

Me: Mhm.

00:38:03

Me: Um, like the risks then, like should have been addressed before those apps are released and not like it's more or less current practice at least in the US to just drop out those applications.

00:38:18

Marloes Veldhuis: Yeah.

00:38:19

Marloes Veldhuis: Really educate people. For instance, just as taking medication. Yeah. You can just buy paracetamol in the in the store and you don't feel pain anymore. But people have to be educated about how how you can safely use it. And I think it also needs to be done with these kinds of applications.

00:38:36

Me: Mhm. Yeah, right. Um. Yeah.

00:38:45

Me: So. Like in general to to summarise, um, would you say it's rather like desirable to implement those technologies or not in your daily practice as a psychologist?

00:39:01

Marloes Veldhuis: I think I would like to see what.

00:39:03

Marloes Veldhuis: I can do, but I think it has to be monitored either by maybe by government. Maybe that would be the nicest so that it won't be capitalised or it won't be. Yeah. To free.

00:39:17

Me: Yeah. So that there's also like human involvement still and they're not like fully automated and.

00:39:24

Marloes Veldhuis: No. Yeah. Mm. Okay.

00:39:28

Me: Yeah, I think that was it from my side. Do you have any comments or questions left or any suggestions? Suggestions for my research?

00:39:39

Marloes Veldhuis: Um, yeah. I was just.

00:39:40

Marloes Veldhuis: Thinking about the VR because I think we, I think that that also might have really nice implications. For instance, with exposure therapy, we also have, I think in our building there's also a department for VR. I don't have much experience with it because we have only the glasses, but we don't have an application. So we have to look up videos on YouTube and they're horrible. Uh, but I think that also would have very nice applications if we can use AI in that to make sometimes it's just very hard to, to create exposure exercises, for instance, with people with slight anxiety or something that you're not going to board a flight together with them or with riding a car or. Yeah.

00:40:21

Me: Mm. Yeah. I think that actually working on solutions for that and they actually testing it to implement it also in those exposure therapies. So I think there are also a lot of possibilities. Um, like also the one psychiatrist I talked about, he was doing research in VR with autistic patients who then could like practice interaction with an avatar via VR glasses. Um, which is like, way more realistic than those. Yeah. Facial expressions we could mimic. So that's also a possibility for application. And they are using it for like people with dementia. Um, just to. Yeah. Get them some activity, like letting them sit in a train if they if that was what they were enjoying beforehand. But you couldn't place them in a train now, so at least they could like relive this experience and are like, yeah, have something to do while the, the caretakers can do something else instead and don't have to fear that the person is running away again.

00:41:32

Marloes Veldhuis: Mhm. That's very nice.

00:41:34

Me: Yeah. They're actually like quite a lot of possibilities but still there are too many concerns that it, it's actually implemented because so many ethical issues are out there.

00:41:47

Marloes Veldhuis: Mhm. Yeah. I can imagine.

00:41:50

Marloes Veldhuis: Your research is really interesting. If you finish your paper, I would love to read it. And yeah, would love to read. Yeah. Know what you found?

00:41:58

Me: Yeah, sure. I can send it to you when I'm finished. I hope it will be end of this or next month.

00:42:05

Marloes Veldhuis: Okay. You know, Good luck.

00:42:09

Me: Yeah. Thank you. And thanks for for joining me with this.

00:43:01

Marloes Veldhuis: Good luck. And yeah. Also like the, like the interview. It's nice to, to learn more about this I think.

00:43:06

Me: Yeah. Okay, great. Thanks for your time.

00:43:10

Marloes Veldhuis: Yeah, you're welcome.

00:43:12

Marloes Veldhuis: Have a nice day.

00:43:13

Me: Thank you. You too. Bye bye.