

Using Artificial Intelligence To Support Students In Improving Their Mental Health

Rosa van Ree

s4631889 - rosa.vanree@ru.nl

Artificial Intelligence, Faculty of Social Sciences, Radboud University – Nijmegen

Master Thesis

Supervisor: Professor Pim Haselager

Second reader: Lotte van Elteren

05-07-2024

Abstract

The number of young students struggling with mental health (MH) issues has reached alarming levels. Carefully designed and implemented, Artificial Intelligence (AI) has the potential to help diminish this problem. This paper explores the role AI can play in ethically and responsibly supporting students in improving their mental health before human help is sought or available. We aim to pave the way for systems that are not mere technofixes but empower users by giving them autonomy in managing their mental health. Using a patient-forward approach, students' needs and wishes from MH-supporting systems were investigated, as well as their experiences with existing technologies. Our findings indicate that successful AI implementation in mental health care must meet the following requirements: It must be helpful, engaging, personalized to the user's care needs and background, trustworthy, technically robust, scientifically valid, fitting into the user's daily routine, accessible, and lawful. This paper also presents a concrete implementation plan, based on state-of-the-art AI techniques and methods, considering their limitations, that could function as a stepping-stone to practically implementing a minimum viable product. The proposed system includes core functionalities like symptom tracking, educational content and exercise recommendations, notifications, a chatbot, personalization options, and crisis-intervention functionality. This thesis challenges the assumption that AI cannot have a place in mental health care. Rather, it shows that to play a vital role in mental health care, implementing AI requires thoughtful attention and should focus on empowering students to take control of their mental well-being.

Using Artificial Intelligence To Support Students In Improving Their Mental Health

Over the last few decades, the problem of young adults struggling with mental disorders (MDs) has taken alarming proportions. In a study done in 2016, the World Health Organization (WHO) showed that over 20% of students fulfill the criteria of being diagnosed with a mental health (MH) condition. However, of those in need, 84% did not receive any professional treatment (Auerbach et al., 2016). Without receiving the needed help, mental health issues often greatly impact an individual's current life, future, and our society as a whole. It is clear that this problem deserves our immediate attention and response.

Artificial Intelligence (AI) has the potential to contribute to diminishing this problem in various ways, but it might also create new additional problems. The goal of this paper is to explore whether, and if so when and how, AI can have a role in supporting students with mental health issues by answering the following question:

How can AI best support students in improving their mental health, before human help is available?

This paper aims to answer the research question using a “patient-forward” approach where we theorize about the role of AI concerning this problem, investigate the end-users' needs and wishes, their experiences with existing mental health technologies, and explore how various AI techniques could contribute to diminishing this problem. Weiner et al. (2013) showed the importance of a focus on “patients' needs and circumstances when planning care is associated with improved health care outcomes”. In a patient-forward manner, the students' needs, wishes, and experiences will therefore be the center of attention used in this paper.

There are two sides to this research question: ‘AI’ and ‘the students’. They are connected by the question of how the first can “best support” the latter. To discover this relation, we start *part one* of this thesis by conducting a theoretical exploration by answering four sub-questions related to the two sides (formally introduced below). For side one – the students – we focus on the following questions: What are their mental health problems and how do those impact their lives? What help could be offered to them before they receive human treatment? Secondly, we must learn about the students' wishes regarding technological support: What do they (not) want from AI? What aspects of existing technologies are helpful and how could these be made more effective? The other two sub-questions connect AI to the context of supporting students in improving their mental health. There are many ways and instances in which to employ AI: It could contribute to different parts of the therapeutic trajectory, support different kinds of stakeholders, and fulfill

different roles when doing so. Therefore, we firstly reason about the function and strengths of AI in this paper's context, so that it can fulfill its role effectively and responsibly. Next, we reason about which AI-techniques could help to play that role and address the needs and wants of the students. When answering these four questions we will constantly focus on the potential advantages and disadvantages that AI might bring. With every opportunity we find we analyze the connected risks or weaknesses and how they can be mitigated.

The answers to our sub-questions guide us to *part two* of this thesis, where we define the relation of "best support" by translating the research question into: "*What design requirements can be formulated to create an AI-based system functioning to support students in improving their mental health, taking into account the lessons from part one?*" The presented requirements - combined with recommendations on how to realize them and a practical implementation plan - show us the answer to our research question derived from the patient-forward path we followed.

In summary, part one of this thesis contributes to answering the research question by discussing the questions presented below, while part two presents a plan to implement the answer.

Part one

1. Why should and how could we use artificial intelligence to improve the mental health of students?
2. Who are the end-users and what is the effect of their mental problems on their well-being and general functioning?
3. What are the needs and wants of students regarding an AI-based mental support system?
4. What technical possibilities does AI offer to realize the needs and wishes of the end-users?
5. Dream scenario: A potential app that meets the end-users' needs and wishes

Ultimately, this research will help us create an overview of aspects required for the successful and effective implementation of AI-based mental health systems, including recommendations on how to meet the requirements. After that, a concrete implementation plan is provided, illustrating how AI could be implemented in a responsible, ethical manner, for it to be most effective in reaching the purpose of its users. This implementation plan also includes a practical work plan, as well as a discussion of technical limitations and gaps in need of further research to bring the implementation to a higher level. The thesis will round up with a reflection on the (un)desirability of personalizable mental health support systems, and a quick overview of the main lessons learned.

Part two

1. Requirements for effective implementation of technology for mental health support
2. Recommendations on achieving the requirements
3. Implementation plan: creating a core version
 - a. Proposed primary functionalities
 - b. Ana again
 - c. Feasibility analysis – a discussion
 - d. Work plan: step-by-step implementation suggestions
4. Conclusion

Knowledge gap

It has been shown that existing technologies created to improve the mental health of young adults are often only minimally (if at all) effective in diminishing the problems of students (Garrido et al., 2019; Grist et al., 2018; Lattie et al., 2019; Tal & Torous, 2017; Vo et al., 2019). However, its potential is clear, as digital MH systems can provide personalized, ever-accessible support during times when demand exceeds supply. Research has been done before on the effectiveness of mental-health supporting technologies, to discover the facilitators and barriers contributing to their use and disuse. This research includes stakeholder perspectives, usability evaluations, and broadscale systemic reviews of the effectiveness of digital mental health interventions (DMHIs). Despite the effort to understand DMHI use and effectiveness, little of these learnings can be found to be applied in actual systems. Generally speaking, there is a risk that research teams working on building DMHIs focus more on implementation issues and less on psychological and user experience (UX) research. Technological enthusiasm or mere “ethics washing” during creation, may cause their products to be suboptimal in practice, or even lead to (intentionally) producing technofixes (Holzmeyer, 2021, Ryan et al., 2019). All in all, there is a gap between technological implementation, psychological research, and ethical desirability. This paper aims to avoid this and bridge the gap between the two by combining patient-forward research literature with a critical examination of AI’s possibilities, to present requirements, recommendations, and a practical implementation plan for the creation of a product that could provide practical, useful support to students to improve their mental health while keeping their needs and wishes primary.

Preliminaries: Outlining basic project choices

Before going into the specifics, some project choices will need to be briefly explained. Firstly, consider the use of the terms “mental disorders (MDs)” and “mental health

issues/problems”. As stated earlier, 20% of all students could be diagnosed with a mental disorder. Different students experience different levels of problems. Before the problems can be classified as a disorder, students might already experience a negative effect on their mental state and life. Despite only about 3% of all students receiving help and therefore having an officially labeled disorder, this paper intends to create support for all students in need of help: diagnosed, undiagnosed but meeting the criteria, or just struggling. The proposed AI-based system can either offer support for those with an MD and mitigate their problems, or prevent incipient problems from escalating. In this paper, we use the terms “disorder” and “issue/problem” interchangeably. However, not all mental health problems can or should be classified as a disorder.

Further, in this project, we decided to limit our research to studies on patient perspectives, selecting a patient-forward approach. Of course, to make the project more dimensional, adding the perspective of mental health experts and clinicians could be a good attribution. Mental health experts often know more about the needs of those with problems, than the patients themselves. Hence, it is important to note that future research incorporating expert knowledge could bring more depth to the analysis and AI requirements.

Lastly, an important theme that we decided not to touch on is the societal debate surrounding technologies like the one in this paper. Despite acknowledging that public acceptance strongly affects the ultimate success of an AI-based system (Apolinário-Hagen et al., 2017), adding that dimension would take away from the core focus of the users’ needs and opinions, which is already a large and complicated topic. After assessing and analyzing the use of AI in mental health care (MHC) from a user-forward perspective, the issue of public acceptance can be addressed by future research.

As stated earlier, this thesis will be divided into two parts: a “Theoretical Exploration” and a “Practical Roadmap”. In the first part, we will theorize about and investigate the subject from a patient-forward perspective, by trying to answer four sub-questions. After, the idea for a potential “dream application” is presented, combining all our learnings into a system that embodies our aim. The acquired insights and the dream app will shape part two, where we introduce a set of requirements for implementation, make recommendations, and present a practical implementation plan. Part two reflects the opportunities, strengths, and weaknesses of AI-based mental health support, addresses practical and technical bottlenecks identified while analyzing our fictive application, and presents recommendations to overcome them.

Part One: Theoretical Exploration

Using a patient-forward approach to analyze the topic, the theoretical exploration is shaped by answering the following sub-questions:

1. Why should and how could we use artificial intelligence to improve the mental health of students?
2. Who are the end-users and what is the effect of their mental problems on their well-being and general functioning?
3. What are the needs and wants of students regarding an AI-based mental support system?
4. What technical possibilities does AI offer to realize the needs and wishes of the end-users?

With the insights from answering these questions - ideas on how we could and should use AI, who the end-users are, what they need and wish for, and what is technically possible - a potential dream application is presented. Importantly, it will be indicated how such an application uses artificial intelligence ethically and responsibly, to realize the users' goal of improving their mental health.

1. Why should and how could we use artificial intelligence to improve the mental health of students?

It is estimated that only 16% of students who suffer from a mental disorder (MD) receive human help (Auerbach et al., 2016). That low number can be attributed to the stigmas surrounding MDs, privacy concerns, a lack of financial means, a perceived lack of time, or simply because students do not recognize their symptoms as problematic or downplay them (Pedrelli et al., 2014). All these reasons prevent students from reaching out for help. The students who do want help from a human clinician face a different problem: long waiting lists due to immense stress on the general mental health care (MHC) system, where there are simply not enough clinicians to treat everyone (Sholevar et al., 2017)¹. The pressure on our MHC system will only grow bigger if the cycle continues like this. It is known that early identification and treatment of MDs leads to a quicker recovery (Pedrelli et al., 2014). Now that more students are waitlisted, their eventual treatment will likely take longer, leading to even more students on waitlists. It is also known that the mental health problems of students often persist for several years, putting even more pressure on the MHC

¹ Sholevar et al., (2017) show a shortage of mental health care providers in the US, however, there are examples of many other countries that deal with similar problems, such as in [the Netherlands](#).

system and the waitlists (Pedrelli et al., 2014). Therefore, we face a serious risk of spiraling down even further.

The existence of technology and AI does not mean using it is our best and only solution. There are hundreds of human-powered projects battling this issue. Also, there is a fair chance that digital mental health care can worsen a student's situation if implemented carelessly, with limited clinical governance, or that an increase in "technofixes" will lead to a reduced investment in human care. Yet, if implemented ethically and carefully, and in balance with human support, AI holds the potential to be more than a technofix. Its strengths lie in its ability to learn from input data, making personalized care on an individual level possible (S. Graham et al., 2019; Lee et al., 2021; Lovejoy, 2019). Because of this, AI has the ability to bridge cultural gaps present in current mental health care technology, and provide support for marginalized groups (Graham et al., 2020; Lee et al., 2021; Schueller et al., 2019). Through analyzing user input data, AI can also accurately predict risk situations, and can therefore intervene at the right moment. Further, it is always available to help a student and analyze their mental state in many instances during the day if desired (Lovejoy, 2019).

In the realm of digital mental health care (DMHC), there are many ways for AI to offer support (Lovejoy, 2019). AI could take over different parts of the therapeutic trajectory, by for example creating a tool that helps with the diagnosis, or a tool used during a user's self-treatment. It could also support other kinds of stakeholders, for example by lowering the workload of clinicians such that they can put more focus on the patients or even create space for extra patients (Lovejoy, 2019; Pokhrel et al., 2021). Consequently, AI could reduce the current stress on the system by improving the quality of care and shortening waiting times.

How you employ AI determines whether you are just *technofixing* or constructively contributing to solving a problem. An AI-based MHC system could have an additional supporting role to one's social system and professional help, rather than replacing it. Current digital MH interventions have not proven to be more effective than human treatment (Garrido et al., 2019), but they can make a positive change during the period before human help is available as shown by Grossard et al. (2017). Such a system could play many different roles depending on the user's needs: from helping a person to obtain a better insight into their situation to giving empathic support, to functioning as an (anonymous) place to reflect on, vent, and communicate about one's problems and mental state. The AI's role could be temporary and fill in the support gap that a user has before professional (human) treatment is available. Once help is available, the system's support might become superfluous or play a smaller role in the user's life. This also mitigates the risk of being a distraction between the student and health care provider (Pokhrel et al., 2021).

Another concern related to such MHC systems is that they negatively affect a person's right to mental integrity (MI). This involves their right to autonomy; to be in control of their mental state and their right against “non-consensual interference with the mind” (Douglas & Forsberg, 2021). The argument is that one’s MI will be compromised by potential privacy issues, emotional dependence, and third-party manipulation of a user’s thinking that these systems could bring. However, one can also argue that systems focused on supporting users to manage and improve their own mental health, contribute positively to MI rights. They bring a right to interventions – similar to a right to human treatment – that “restores and sustains mental functioning and promotes its development” (Paz, 2024). MH supporting systems could even increase one’s autonomy over their mental state, as long as it is conducive to self-determination. It is a mistake to view digital mental health systems as a malicious replacement for human care when they bring the opportunity to self-administer well-needed, additional support.

Of course, we need to be cautious about the functioning of the potential AI-based MH systems intended to be deployed. First, creating a system that pushes the user or forces thinking patterns or ideas on them must be avoided. The system *should not* take away autonomy, manipulate, or make its users dependent. On the other hand, it *should* support the user in learning and gaining insights about themselves, and give them tools to self-manage their growth. The system also *should* be transparent about its workings and outputs, for the user to be aware of the system's role in their developmental process.

Above all, a clear difference needs to be kept in mind between: “*Using AI to improve the mental health of students*” and “*Using AI to support students in improving their own mental health*”. Focusing on the latter aim will help us maintain a supportive, tool-like intent, allowing AI to contribute responsibly.

2. Who are the end-users and what is the effect of their mental problems on their well-being and general functioning?

When researching how AI can best support our end users in helping them improve their mental health, it is important to understand their mental disorders (MDs) or issues, and situations better. What MDs are most common and what are their characteristics? How do they develop and how are they treated? Gaining a deeper understanding of the end-users will help us help them better.

It has already been shown that there are numerous reasons why students in need of help do not receive any human treatment, with all its consequences. There are various apps on the market trying to improve the mental health of this group, however, they do not seem to be effective and are currently underused (Vo et al., 2019). To illustrate students' mental health problems and how

they negatively impact their lives, I present a fictive case scenario², describing “Ana’s” situation. The issue presented in this scenario is based on one of the most common mental health disorders found among students and exemplifies corresponding symptoms. The scenario also shows why professional, psychological, help is not received or sought for and shows us narratively the negative impact mental disorders can have if the current system is not changed in one way or another. Further, it shows how Ana tries to use the current most accessible implementations of technology for mental illness support.

As you read this thesis, we encourage you to consider Ana as a representation of (a large, representative segment of) the 'end-users'. However, recall that not every intended end-user will have a fully developed mental disorder. Students struggling with all kinds of (levels of) mental problems should benefit from a proposed AI-based support system, especially when they do not receive help from a human therapist (yet).

Case scenario

More often than not, Ana feels sad, empty, and a little hopeless. Her daily activities, from taking classes to seeing her friends, do not interest her anymore. She has noticed weight loss, is often fatigued, and lacks the concentration to study. On top of that she experiences worthlessness and feelings of guilt toward her family. As a result, Ana has been failing most of her classes and is slowly losing all contact with friends and classmates.

Though she suspects that she might suffer from depression, she does not dare to talk to anyone about it. In her culture, seeking treatment for mental health problems indicates craziness and would bring shame to her family. It would mean she has failed as a person. As a result, she tries to hide her state as best as she can, lies to her family and friends, and searches for alternative ways to battle the feelings she has. She has installed a couple of apps, which taught her that building healthy habits can help in battling depression. However, she is skeptical that the simple features of an app could actually improve her state. Likely due to her depression, she also lacks the motivation to keep engaged enough with the apps to properly develop those habits.

Further, when using the different apps, she’s found herself frustrated that the contents of the app do not match her cultural background and understand her problems. The customization options don’t match her needs, as there is nothing in the apps to help her fight her feelings of guilt and shame. She has also noticed that some of her harmful, suicidal thoughts only worsen during or after the exercises and meditation sessions that the apps encourage. This has made her scared to

² Inspired by the work of Vlek et al. (2012)

use the apps altogether and respond to their notifications. Every so often, she installs a new app, but it typically only gets used for a day or 3.

A mental disorder is defined as: “*A syndrome characterized by clinically significant disturbance in an individual’s cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning. Mental disorders are usually associated with significant distress or disability in social, occupational, or other important activities.*” (American Psychiatric Association., 2013).

The case scenario above presents a student suffering from (the onset of) a classic major depressive disorder. Together with anxiety disorders, mood disorders (like this one) are the most common among students (Auerbach et al., 2016). Mood disorders occur in different forms, with other types of symptoms and varying treatment needs. Ana has a classical presentation of depression, however depression with more atypical symptoms presents itself more frequently. It is not unusual anymore for people suffering from depression to brighten up while doing something that makes them happy. Anxiety disorders occur in many forms, from specific fears to social anxiety to panic disorder. Symptoms associated with anxiety disorders are disproportionalities of stress, worry, and fear. Sometimes a student’s disorder is caused by excessive substance use, during or after intoxication (American Psychiatric Association., 2013). Not every student will meet the full criteria to be diagnosed with a mental disorder, however, manifestations of some symptoms are known to many students.

There is no unambiguous answer as to why so many students develop MDs and why they generally persist longer (Pedrelli et al., 2014). Often, an interplay of factors contributes to the development of an MD. Think of genetics, academic pressure, social pressure, financial stress, and the transition into adulthood (American Psychiatric Association., 2013; Ibrahim et al., 2022).

Untreated or unsupported, those experiencing mental health difficulties often develop unhealthy coping mechanisms consequently worsening their disorder, and increasing their eventual treatment duration. Having helpful support while waiting for human treatment can therefore be of value to the general mental health care system. This support, potentially digital, could learn from established treatment techniques, to increase their quality and effectiveness (Grist et al., 2018; Pedrelli et al., 2014). Favored techniques for mood and anxiety disorders include behavioral therapy, cognitive therapy, or cognitive-behavioral therapy. In these types of therapy, there is a focus on the relationship between behaviors causing difficulties and unhelpful, pessimistic thinking. Therapist and their patients try to disrupt or change negative behavioral patterns or thoughts (*Depression Treatments for Adults*, n.d.). Further, a digital MH-support system could learn from principles and techniques derived from interpersonal psychotherapy, mindfulness,

supportive therapy, or exposure therapy (*Depression Treatments for Adults*, n.d.). All including helpful treatment techniques that can aid different students with different needs.

3. What are the needs and wants of students regarding an AI-based mental support system?

Within the field of mental health support, researchers and clinicians have been (trying to) use technology to improve the mental health of students for a while now. Serious games, virtual reality (VR) applications, internet-based cognitive behavior therapy, and mental health care (MHC) chatbots are all examples of previously created digital mental health interventions (DMHIs) (Apolinário-Hagen et al., 2017; D'Alfonso et al., 2017; Fitzpatrick et al., 2017; Grossard et al., 2017; Lattie et al., 2019; Prakash & Das, 2020). Yet, the most common occurrence of digital MHC are “apps”; self-help, lifestyle apps that focus on creating healthy habits, understanding one’s symptoms, using check-in moments, sometimes providing exercises, and a form of community. The existing apps use tracking functions and give some statistics about one’s symptoms. Even though over 1000 slightly varying apps exist, indicating user interest (Tal & Torous, 2017), they are severely underused and not proven effective in improving students’ mental health (Garrido et al., 2019; Grist et al., 2018; Lattie et al., 2019; Tal & Torous, 2017; Vo et al., 2019). The failure of many apps can be attributed to the fact that they do not fully match the wants and needs of the end-users and lack the factors contributing to the usage of DMHIs as specified below. Yet, the use of apps to support students in improving their mental health is a promising one, due to their broad accessibility, availability, and the end-users’ familiarity with such technology (Inal et al., 2020). We therefore choose to focus on AI-based solutions, implemented in applications for mobile phones, while taking into account insights from research on DMHI usage from a patient perspective.

When wanting to create an application that supports students successfully in improving their mental health, we must start by studying their current attitude towards MH apps and their *needs* and *wants* for future products.

Factors related to the use and disuse of digital mental health interventions

It has been shown that possible users of DMHIs see their potential, but still have low intentions of using them (Apolinário-Hagen et al., 2017). Perceived usefulness and helpfulness in improving the user’s situation logically *increase* DMH-system usage (Abd-Alrazaq et al., 2021; Apolinário-Hagen et al., 2017; Prakash & Das, 2020). A student's current level of stress or need for help increases usage as well (Borghouts et al., 2021). Further, also positively associated with DMH-app use are social influences, and having used services in the past (Borghouts et al., 2021).

Lastly, engaging patients in managing their own healthcare plan and increasing their empowerment is strongly associated with the uptake of apps (Borghouts et al., 2021a; Vo et al., 2019).

However, there are also several factors contributing to the *disuse of DMHS*, such as a low level of trust in the system, provider, and/or technology in general (Vo et al., 2019). The distrust is related to safety concerns, including privacy concerns and fears about worsening their health situation, as well as a lack of research evidence for the effectiveness of apps (Borghouts et al., 2021; Graham et al., 2020; Prakash & Das, 2020). Users are also held back by stigmas surrounding mental health issues (Graham et al., 2020). However, if they do decide to try an app, they find that many systems are not appropriate concerning their background, care planning wishes, and clinical picture. All in all, there is a lack of personalization options, which could have contributed to the system's appropriateness (Borghouts et al., 2021a; Vo et al., 2019). Also, often, using the app does not fit into users' daily routines as apps can be too demanding, especially for someone suffering from MH problems (Graham et al., 2020). Lastly, a big factor is the accessibility of the system, e.g. its cost, but also awareness of the existence of the service (Apolinário-Hagen et al., 2017; Borghouts et al., 2021; Graham et al., 2020; Vo et al., 2019).

From the users' perspective: values, wants, and needs

As expressed by users in prior research (Borghouts et al., 2021a; Tal & Torous, 2017; Vo et al., 2019), all features of the app should contribute to (or at the minimum not obstruct) *user empowerment*. The system should empower its users by giving them the tools to gain more *control* and *autonomy* over their mental well-being. Users should be *self-determining* about when and how they want to be helped by an AI-based MH-support system. Further, using the app should result in a raised *self-esteem* and a sense of doing better.

How do we make this requirement more concrete? What features do we need to design to fulfill the needs of empowerment, control, and autonomy?

It is also important to note that “*not one specific method fits the needs of all individuals, however, there is also not one specific method that fits the needs of one individual at all times*”. If it was that easy, the problem would not be as big. When designing an AI-based system to support the mental health of many different people, there needs to be awareness of the differences between those people and create something adaptable that can support many. Despite this, research has shown that there are specific features that can optimize the effectiveness of such a system for many different users (Abd-Alrazaq et al., 2021; Borghouts et al., 2021; Borghouts et al., 2021a; Garrido et al., 2019; Graham et al., 2020; Hudson et al., 2022; Oyeboode et al., 2020; Pokhrel et al., 2021; Prakash & Das, 2020; Schueller et al., 2019; Titov et al., 2019; Vo et al., 2019). These will be divided into “*user wants*” and “*user needs*”. The ‘wants’ represent a set of features,

functionalities, or characteristics wished for by users in a digital support application. ‘Needs’ on the other hand, are matters proven relevant to users to be able to get better and improve their mental health. The users did not wish for them specifically – like they do with the wants - but they are needed to be included to support them better.

User wants

1. The ability to track, log, and “measure” one's mental health, using assessment tools or quizzes. Additionally getting insight into the results, getting insights from the AI system, and using the system to reflect on those insights (Borghouts et al., 2021a; Hudson et al., 2022; Oyebode et al., 2020).
2. Educational content and exercises: Inclusion of mental health information on disorders, diagnosis, treatment, therapeutic techniques, statistics, etc., and practical exercises that implement the learnings and help users cope with their symptoms (Hudson et al., 2022; Titov et al., 2019).
3. The option to personalize the system; match the user’s clinical picture, culture, background, wishes about app features, and planning care (Hudson et al., 2022; Oyebode et al., 2020; Vo et al., 2019). Further, data points approved by the user should drive content recommendations. A feedback feature to directly adapt the personalized profile could be included. Additionally, a potential chatbot included in the app should also be tuned to individual treatment recommendations (Abd-Alrazaq et al., 2021).
4. Virtual rewards, related to in-app goal setting, and gamification features (Garrido et al., 2019; Oyebode et al., 2020).
5. Social connectedness; meaning a social aspect of some form: e.g. forums, messaging functions, social circles, peer-to-peer counseling, etc. (Borghouts et al., 2021a; Graham et al., 2020; Oyebode et al., 2020; Schueller et al., 2019; Titov et al., 2019; Vo et al., 2019).
6. A form of human support; being able to connect with a human therapist or have health worker supervision (Pokhrel et al., 2021; Vo et al., 2019).
7. The system should fit into the daily life of users; using the system should take little effort to remain accessible and engaging (Graham et al., 2020; Hudson et al., 2022).
8. Enjoyable to use; hedonic motivation is caused by usage (Abd-Alrazaq et al., 2021; Oyebode et al., 2020; Prakash & Das, 2020).
9. Reminders and notifications (when necessary and dependent on the user’s preference) (Hudson et al., 2022; Oyebode et al., 2020).
10. A crisis functionality; A panic button for users or intervention by the app at a user’s low point (Titov et al., 2019).

11. High-quality content that is: relatable, interactive, relevant, appropriate, stigma-free, varied, personalized, and valid/credible (Abd-Alrazaq et al., 2021; Garrido et al., 2019; Hudson et al., 2022; Oyeboode et al., 2020; Schueller et al., 2019; Vo et al., 2019).
12. Anthropomorphized system characteristics, while remaining identifiable as an AI, such as personality, perceived empathy (sympathy, care, warmth), and intelligence (e.g. sufficient linguistic skills to understand the user's input and produce appropriate responses (Abd-Alrazaq et al., 2021)) (Prakash & Das, 2020).
13. To have trust in the system and its content. Additionally, trust in the provider to not bring harm, respect privacy concerns, and handle the users' data with care (Abd-Alrazaq et al., 2021; Graham et al., 2020, Oyeboode et al., 2020; Prakash & Das, 2020; Pokhrel et al., 2021; Vo et al., 2019).
14. Using the app should be free or have minimal cost to enhance accessibility. (Hudson et al., 2022; Oyeboode et al., 2020; Vo et al., 2019; Prakash & Das, 2020).
15. The app should be technically stable, run smoothly, and function without flaws (Oyeboode et al., 2020).

User needs

1. Person-centered shared decision-making: Involving patients in the planning of their own health care has been shown to improve the effectiveness of treatment (Tal & Torous, 2017; Vo et al., 2019; Weiner et al., 2013).
2. Normalizing MH problems and treatment: Acceptance of oneself, including accepting one's mental disorder, results in more willingness to seek help from a therapist or one's social system, which will subsequently lead to faster and better recovery (Pedrelli et al., 2014).
3. Use of acknowledged therapy forms and techniques (Naslund et al., 2017; Pedrelli et al., 2014), such as cognitive behavioral therapy (CBT), and providing research-based content. Using the right therapy form that fits the individual and their clinical picture leads to a safer and better treatment process and recovery.

4. What technical possibilities does AI offer to realize the needs and wishes of the end-users?

Before determining how AI can contribute to realizing the needs and wishes of the target group, we will first explore AI's possibilities regarding mental health support without limiting ourselves. Exploring the various methods and techniques, without setting any ethical boundaries or raising performance questions, will be a helpful exercise. Firstly, it will stir our creativity in

finding possible AI solutions to reach the goal of improving the mental health of students. Secondly, it will make us aware of the potential threats and weaknesses that accompany certain directions of employing AI for MHC and allow us to be cautious and minimize the risks.

Possibilities of AI

Artificial intelligence models use lots of data to find patterns in them, and classify or make predictions about new data additions. The models can output all kinds of insights about the input data, such as recommendations, labels, predictions, or relations between data points. Students coping with MDs have many personal data points that could be used as input data, to produce statistical insights about themselves, in the form of patterns or predictions. For example, students could track their sleeping hours, physical activity, and their mood, to try to find a relation. Using AI, patterns between data points can be found that we might never suspect are connected or influence a person's well-being. For example, a changed typing behavior (e.g. less interpunction and shorter sentences) could indicate a new depressive episode. Predicting MDs like this could help us intervene quicker and make us more aware of someone's mental state and well-being. Table 1. shows us various examples of types of personal data that could be tracked (Place et al., 2017; Sano et al., 2018; Tal & Torous, 2017).

Apart from using different types of data, different types of AI models can be implemented as well. To simplify the realm of AI, we will only differentiate between supervised and unsupervised machine learning (ML) models for now. In *supervised* ML, the input data is labeled, and it is known which output targets to expect. This type of ML could be useful if we want to diagnose a user based on their personal input data. In *unsupervised* ML we are trying to gain insights without any guidance and pre-existing bias (apart from bias present in the dataset). By doing this, unexpected relations between the user's data points can for example be found, possibly giving the users new valuable insights specific to their mental health.

Acknowledged AI methods and techniques (such as linear regression or hierarchical clustering), are being applied to achieve different technological abilities. Think, for example, of *image recognition*; various techniques could be used to achieve this, from using convolutional neural networks to feature extraction. These abilities and general techniques could be applied to achieve the users' product wishes. We could for example perform content or sentiment analysis on clinician and user notes, to find hidden information about how the user is feeling from written text. Further, computer vision techniques for emotion recognition and image analysis could be used to monitor the user through their camera. We could also use content generation methods, such as generative adversarial networks (GANs), to produce images or music with a calming purpose, to help cope the users in difficult situations. Large Language Models (LLM) can help us create

sophisticated chatbots capable of acting like a therapist, using specific therapy strategies or psychological techniques. All of these models combined could make for personalized AI support that can act on the model's insights, recommendations, and predictions, listen to its users, and aid them in all sorts of tasks.

Despite being set on creating a mobile application, an exploration of the use of AI in combination with other technology and hardware could be helpful to discover even more possibilities. Think for example of virtual reality (VR) and augmented reality (AR) applications. Using a combination of generative networks, and computer vision techniques for image recognition and object tracking, we could manipulate the user's surroundings when they use VR glasses. By doing this, we could help the user deal with panic attacks or depressive episodes by trying to make them calmer and happier. VR and AR could also be used for exposure therapy, where the user can train fearful situations. These VR/AR ideas could possibly be implemented into a mobile application as well. Another idea could be to implement AI models into robotic hardware to - to give some examples - imitate a human psychologist (with a physical presence), build on the idea of a robotic support animal, or provide physical help and stimulation for doing daily chores or other activities that might take too much effort for people suffering from mood disorders.

Risks and how to mitigate them

All these applications of AI come with weaknesses and risks. It is important to be aware of those to prevent harm from happening. We have seen multiple applications of AI that can be invasive to a user's mental integrity and can lead to a *big brother* effect. Tracking many very personal data points, and analyzing written text or facial expressions form a risk to the human right to privacy, and therefore to the user's mental well-being. Another weakness of drawing conclusions from user data is *bias*. There is a high chance that the model's findings would be generalized, which increases the risk of misdiagnosing users and exposing them to irrelevant, possibly harmful content. Next, there is a need for sufficient human oversight when letting AI generate content or apply risky therapeutic techniques. Humans should remain in-and-on the loop to make sure the AI does not cause its users harm. Further, the ethical desirability of manipulating and lying to the end-users should be questioned. Chatbots or robots that pretend to be human therapists, or are undistinguishable from one, could form a threat to the integrity of MHC and disrupt the user's trust. Lastly, the disadvantages of using other types of hardware over mobile phones need to be considered. Despite the functionalities that they bring, VR glasses and robots are less accessible – as they are expensive - and less usable on a day-to-day basis, as you are not carrying them around at all times when you need support, as most students would with a phone.

The ethical and safety risks, show us where measures are needed to mitigate harm. For example, a (mental) big brother effect and feelings of invasion, could be tackled by providing the user with more agency and autonomy over what is and isn't tracked and at what level. The model's bias and its (harmful) consequences can be mitigated by bias measures such as data quality management (making sure training data is diverse and representative), implementing bias detection algorithms, and regular model evaluation by experts. Potential generated content should be monitored and assessed by mental health experts as well. Further, it is generally questionable whether AI systems should be given the responsibility of a sensitive task like practicing therapy. Similarly, there is a need to prevent the delusion of users and ensure a clear distinction between technology and humans. A user should always know they are interacting with a machine and be aware of how that differs from human emotional support and expertise.

In this part, some of the potential of applying AI to the problem of a rise in mental disorders has been shown, as well as the risks that are attached to those possibilities. In the next section, the ideas presented here will be incorporated into a fictive mobile application, to demonstrate how a certain app could function for potential end-users.

Data type	Example data points
Physiological data	Heart rate, blood pressure, internal temperature
Geographical data	GPS, accelerometer (indicating pace), time spent in specific places (such as one's bed, shower, in or outside the house)
Data regarding activities	Amount/time spent on physical activity/exercise, work, household activities, sleep
Specific social behavior	Social media usage, calling, texting, meeting with friends, conversing with strangers
Consumption data	Intake of food, water, and alcohol: how much and when.
Mobile phone behavior	App usage, internet usage, typing behavior
Audio and video data	Microphone and camera monitoring, image/video analysis
Self-assessment	Mood-trackers, questionnaires
External factors	World events, weather

Table 1. Potential input data (personal to the user) that can help us gain insight into one's mental health.

5. Dream scenario: A potential app that meets the end-users' needs and wishes

To conclude part one of this thesis, our learnings and insights will be combined into a “*dream app*” for Ana and the other end-users. The fictive app fulfills an additional role in a student’s support system before professional help is sought or available. It uses AI’s technical possibilities, to realize their needs and wants responsibly, safely, and in the users’ best interests. Using the app, will empower the students, and give them more control and autonomy in managing their mental health. The app presented below ties the research done in this first part to the case of Ana. This thought experiment aims to design something more effective and successful in helping her, compared to apps she used before.

Ana has installed a new app that has been funded by the European Union to support students in self-managing their mental health and is free to use. The EU project group has tried to stick as closely to the EU Guidelines for Trustworthy AI and General Data Protection Regulation (GDPR) as possible and tries to show that within the app. Also noticeable, is stakeholder and expert inclusion: From expert comments and valid sources provided to increase the quality and strengthen the trust in information in the app to a separate page where you can find an overview of user input, how the input is implemented, and a function to contribute yourself. Consent is presented playfully in a gamified way and is always adaptable throughout the use of the app (e.g. delete or add data). Furthermore, there is a high level of transparency with thorough, easy-to-understand explanations of the workings of the AI models and their generated recommendations and insights.

Overall, the app works as Ana expected and is technically robust. In case of mistakes, there is a function to indicate dissatisfaction. Further, the provided content is credible, as the exercises and techniques are provided by certified psychologists. The content Ana works with allows her to work on healthy coping mechanisms, makes her learn about herself and her issues, and feel better during the day. Through the workings of the AI model, the recommendations are matched to Ana’s clinical and emotional needs. Also, the in-app chatbot seems sympathetic, kind, and intelligent. It understands Ana’s inputs on a high level and produces appropriate and useful responses.

The app's main strength lies in personalization functionality. Ana can completely customize her health plan: From which symptoms to track, to what background context to take into consideration. Does she want to gain insights from the AI, or look at the statistics herself? Does she want guidance in reflecting on those insights? Is a form of gamification or goal-setting desired? How about exercises, and how much? What are convenient times to check in? On what does she want the focus of her support to be? Creating healthy habits? Being more peaceful? Ana can set up the app so that it fits into her specific daily life, without being demanding.

All these choices were too overwhelming when she first started using the app. However, the app includes various standard versions, with varying levels of extensiveness and features available. Ana started with a standard version, and after using the app for a bit started to explore different features and change her healthcare plan accordingly.

Whilst using the app, it will get to know Ana better: understand her (cultural) background and what that entails. It develops alongside her app usage and from the in-app feedback system. This leads to better-matched recommendations (e.g. for exercises, articles, or one's care plan). When Ana's needs change, the health plan can be changed accordingly. Further, together with the app, she can reflect on her process and development: What therapeutic techniques work for her, and which topics could use more attention?

Ana does not have to be proactive all the time if she is not capable. For example, the system automatically suggests re-evaluating one's health plan if it notices a diminished interest in the recommended content. However, this is purely to benefit the user. The product owners have no capital interest in keeping their users engaged.

A social aspect can also be found in the app. Users get anonymously matched into little groups, based on similar clinical pictures and contextual factors. The app motivates the group to learn from and with each other: share insights, and tips, and check in together. This feature makes Ana feel less alone, more normal, hopeful, and empowered to get through her problems. If necessary, the app also offers the availability to make an appointment to video call or chat with a human clinician.

Planning one's own care leads to feelings of control and empowerment and creates more user engagement. Self-determining one's care also leads to more trust overall. This app is meant to help people understand that having an MD is nothing to be ashamed of and that there is pride in taking control of your own health and life. With the app, a user can learn about themselves, and their disorder, and learn how to cope more healthily. The small successes the app will stimulate and celebrate can make the user feel stronger, more in control, and more capable of beating the disorder.

In part one of this thesis a theoretical exploration was conducted. As we have seen, mental health issues can greatly impact a student's life, both professionally and socially, next to feeling unwell. Though AI offers the potential to decrease the mental health problems of students, it needs to be employed with care. As shown, there is a place for AI systems in MHC where it has a supportive, tool-like role, and is not used as a technofix replacing human care. Potential users indicated that a system should empower them, and give them more autonomy and control over their mental state. The system should make it easy to work on one's mental health by offering

support and providing relevant content and tools. To achieve this, users expressed a set of wants that their ideal app contains, such as the option to personalize the app, gamified features, and interactive exercises. Certain AI techniques that can help realize these wishes have been explored. For example, the use of a clustering algorithm can help us cluster users together in social circles based on their clinical picture and socio-ethnic background. Another option is to gain insights into the relation between symptoms, moods, and events by using machine learning models that can find hidden patterns. After a fictional dream app was presented that used AI to include most user wishes. Functioning as a support-tool, it empowered Ana to improve her mental health.

Though an app like this sounds promising, it leaves us to wonder whether it is actually possible to create something like this with state-of-the-art technology. Has the field of AI advanced to a level to perform to the standards of the dream application? In part two of this thesis, we will answer this question by providing design requirements for AI-based mental health support systems, to create an application that matches the users' needs and wishes like the dream app does. Additionally, AI's current strengths and weaknesses will be made visible by a practical implementation plan, that will provide a stepping-stone to implementing a first minimum viable product, resembling the dream app. It will be shown what actually can be done with current AI models and techniques, and what remains a dream for now (with the potential to become reality with more research and technological advancement).

Part two: Implementing AI for Mental Health Support – a Practical Roadmap

Let us recall the paper’s research question: our goal is to discover how AI can best support students in improving their mental health before human help is available. In part two of this thesis, we shift from researching the problem and possible solutions from a user perspective to answering the questions by presenting design requirements and a practical implementation plan. The research questions translates into: *“What design requirements can be formulated to create an AI-based system functioning to support students in improving their mental health?”* We aim to combine our findings from part one - as summarized above - into a practical roadmap that can help us create applications or other digital mental health interventions, responsibly and ethically, to help reach the students’ goal of feeling better mentally.

Too often, perhaps due to technological enthusiasm, development teams overlook the needs of their users, address them too late, or misinterpret them. Plenty of research into factors that contribute to the effectiveness of technological use for mental health improvement from a patient perspective exists. However, it often strands after publishing, rather than being picked up by app developers. Part two of this thesis aims to bridge that gap by creating requirements driven by patient-forward research and taking it one step further: We will view our learnings through AI glasses and translate the requirements into a practical implementation plan, ready to be executed.

This brings us to three outputs, that together form a framework for effective implementation of AI for MH support: a set of requirements, recommendations to realize the requirements, and a concrete implementation plan that combines the requirements and recommendations in a project-like manner. The requirements presented below are drafted by combining insights from our prior research and thinking, as shown in the dream applications. They cover a range of subjects: From fulfilling specific user wishes to feasibility issues, to requirements for the successful uptake of an application. In the recommendations section, we advise on overcoming practical bottlenecks and implementing user-focused functionalities safely and responsibly, to fulfill the requirements. After, the implementation plan shows how state-of-the-art AI ideas can be used to realize an application that meets the needs and wishes of the end-users. It also shows where there is a need for further research and development to create the most promising implementation possible.

This thesis will conclude with a reflection on the (un)desirability of personalizable mental health support systems and provide and provide a quick overview of our learnings to answer the research question.

1. Requirements for effective implementation of technology for mental health support

Firstly, nine requirements will be presented drawn from learnings from part one of this thesis. Specifically, from prior research about factors contributing to the use and disuse of existing digital mental health interventions and apps, and user wishes/needs. (Abd-Alrazaq et al., 2021; Apolinário-Hagen et al., 2017; Borghouts et al., 2021a; Garrido et al., 2019; Graham et al., 2020; S. Graham et al., 2019; Hudson et al., 2022; Lattie et al., 2019; Mohr et al., 2017; Oyeboode et al., 2020; Radovic et al., 2022; Tal & Torous, 2017; Titov et al., 2019; Vlek et al., 2012; Vo et al., 2019; Wykes et al., 2019). The points mentioned here are all required to create a product that is effective in reaching the users' goals and also has the ingredients for them to start - and keep - using it. When complying with these requirements the product will be: helpful, engaging, customizable to the user's care needs and background, trustworthy, technically robust, scientifically valid, fitting in the user's daily routine, accessible, and lawful.

1. **App performance and usability:** The first requirement for effectiveness is to create a usable, well-performing product (Abd-Alrazaq et al., 2021; Oyeboode et al., 2020). An app is considered *usable* if it is easy and convenient to use. Additionally, it needs to be *useful*, meaning helpful in reaching its user's goal. Most importantly for an app to be useful, it needs to include functionalities wished for by the users and meet their (and potentially society's) expectations (Titov et al., 2019). Other factors contributing vary from good aesthetics and easy navigation to app-stability and customization options (Garrido et al., 2019; Hudson et al., 2022; Oyeboode et al., 2020). Lastly, for an app to be usable, it is required that it should not cause harm to the user's well-being.
2. **Engagement:** Staying engaged with an app can be a tough assignment when suffering from a mental disorder (Borghouts et al., 2021a). However, an app can only improve a student's situation if they keep engaged with it. Engagement is naturally the result of other requirements, such as accessibility, trust, app performance, and appropriateness. However, aside from executing these other factors well, various aspects can contribute to even more engagement. These aspects include social connectedness, hedonic motivation, and app acceptance. Also, the app should easily fit into the day-to-day of users to increase engagement rates.
3. **Personalizability:** Personalization options are an important user want (Hudson et al., 2022; Vo et al., 2019). However, it is not just "a want" as it contributes to the *appropriateness* of the app for the user and therefore contributes to effectiveness and helpfulness in supporting a user's MH. This feature makes the application suited for all different users and enhances their self-agency and self-determination when they customize

their health plan and goals. Adding this feature also diminishes discrimination against minority users and contributes to being open to more diversity (Graham et al., 2020; Lee et al., 2021; Lovejoy, 2019; Schueller et al., 2019).

4. **Obtaining trustworthiness:** This is a requirement for the effectiveness of an application, as a product that is not trusted will not be used and accepted and therefore has no chance of contributing to the goal of supporting students in improving their MH. Trust also hugely impacts other requirements, such as usability and engagement.

Trust is highly context-dependent; all kinds of factors related to the user, as well as public acceptance of the app, can impact trust. Trust has been found to take place on four levels each of which requires different strategies for obtaining trustworthiness. There can be differentiated between (1) “trust in the provider” (Prakash & Das, 2020; Oyeboode et al., 2020), indicating the distrust of users in the provider handling their data with care and creating well-performing products for them. Further, we find (2) “trust in technology” (Prakash & Das, 2020), indicating the scariness of relying on technology for many people. The last two levels of trust are about (3) “the helpfulness of the application” (Oyeboode et al., 2020), and (4) “the safety of the application” (Prakash & Das, 2020), indicating a wished credibility of the product, where it matches the user’s expectation to deliver a high-quality product, that does not harm their mental well-being.

5. **Technical feasibility and robustness:** The effectiveness and success of a product highly depend on the feasibility of implementing the ideas and the robustness of said implementation (Borghouts et al., 2021a; Oyeboode et al., 2020). This requirement covers all issues related to the technical workings of a system (including all wished-for functionalities); from the feasibility of creating the models, including obtaining high-quality training data and the resulting model’s robustness, to transparency, explainability, data processing, and consent, all the way to the environmental impact of the app.
6. **Creating a scientifically supported product:** Making sure the product is scientifically supported is highly important for its effectiveness as well (Tal & Torous, 2017; Vo et al., 2019). Most importantly, delivering high-quality content contributes to a more appropriate solution that is therefore more likely to support its users’ goals. Further, it contributes to the product's safety level, the user’s trust, and perceived usability and helpfulness, leading therefore to higher effectiveness, engagement, and uptake.
7. **Closing the gap between research settings and in-practice usage:** Many studies have shown that an important factor for the ineffectiveness of existing DMHIs can be traced back to the gap between research settings and clinical practice or the daily lives of users (Graham et al., 2020; S. Graham et al., 2019; Hudson et al., 2022; Mohr et al., 2017; Titov

et al., 2019). Apps that seem promising in laboratory settings or when participants are paid to test them often do not suit the actual lives of the end users. In practice, users find it hard to stay motivated to keep using the product. They find it too much work, it does not fit into their daily routine, or their mental state prevents them from adopting the app.

8. **Accessibility:** As mentioned earlier, poor accessibility negatively impacts the use and uptake of an app (Apolinário-Hagen et al., 2017; Vo et al., 2019); creating an accessible product increases the likelihood of adoption. Further, a societal service product - like an MH-support app – must ethically benefit many users (Vlek et al., 2012; Wykes et al., 2019). There should be no unfairness and discriminative factors (e.g. race or income) in play. Sub-challenges regarding the accessibility of an MH-support app are: (1) Creating awareness of the existence of the product, (2) making the product accessible *in* use (e.g. easy to understand), (3) making the product accessible *to* use (e.g. little/no payment attached).
9. **Legislative requirements/feasibility:** The app must be compliant with national and international law. Not only does this contribute to the user's trust, but law also meets and mitigates many ethical issues and safety concerns for a user's health and human rights (Radovic et al., 2022; Vlek et al., 2012). Compliance with legislation like the GDPR or the European AI Act therefore contributes to a more responsible AI-based product (Wykes et al., 2019).

In extension to these requirements, there is a need for awareness of the implications of the target group's characteristics when designing products. Because of the students' mental disorders, we need to act with more care and keep the students' limitations and vulnerabilities in mind. Their minds are more sensitive to harm, and already being mentally less stable. We must be careful not to worsen their mental health. When providing them with new information and experiences, we must be conscious not to be too forceful or manipulating in any sense and bring them into distress. Further, the students' social, cultural, and professional backgrounds are important to take into account, as those greatly impact the success of the treatment process.

The implications of designing for this target group are relevant to all the requirements above. For example, MH issues often result in a lack of motivation or energy to continue with one's regular life. Asking them to do new things can be paralyzing, making the requirement of engagement harder to fulfill (Borghouts et al., 2021a).

Lastly, remember that mental disorders can lead to different impairments for different people, or in different contexts. A well-designed MH app therefore includes customizability features to match the needs of different users in different situations. For example, one student might need a

decision-aid (as making decisions can be hard due to mental issues), whilst another needs stimulation to make their own decisions to gain back a sense of control.

2. Recommendations on achieving the requirements

To meet the requirements for the effective implementation of systems that support students in managing and improving their mental health, a series of recommendations has been drafted. All requirements mentioned above are connected by the goal of leading to effectiveness, and the recommendations are presented accordingly; from the premise of a connected whole. Often, specific recommendations are relevant to more than one requirement. Therefore, we chose to divide the recommendations into groups, also called “themes”. Per theme, it will be indicated how a recommendation contributes to realizing the requirements. An overview of the recommendations can be found in Table 2. at the end of this section.

Sometimes, fulfilling one specific requirement helps to realize others. For example, when meeting the requirements of *usability* and *trustworthiness*, the product is required to be *technically robust*; so realizing the requirement of technical feasibility contributes to multiple others. In another case, we can see how some requirements are simultaneously recommendations fitted for other requirements. For instance, for the product to be *engaging* it needs to fit into the daily life of its end-users. We however chose not to make a separate recommendation about that, as this point is already addressed when fulfilling the “*closing the gap between research and IRL settings*” requirement.

Primary & secondary functionalities

The recommendations presented below are grouped to help us meet the requirements step-by-step. Firstly, we must focus on implementing a set of primary functionalities, based on the end-users’ needs and wishes. Together, these primary functionalities create a minimum viable product. It is recommended that this product includes logging and tracking functions, educational content, exercises, a light option to customize one’s health plan, reminders and notifications, a crisis function, and a chatbot with varying functionality, e.g. help reflect on insights about the user’s mental state, provide empathy, or help with making decisions. AI’s role in implementing the primary functionalities and the plan’s technical feasibility will be described in more depth in the implementation plan below.

After, a set of secondary functionalities could be implemented. Though less important for a core implementation, these features will help us fulfill even more requirements. First, to enhance the app’s usability and performance, an in-app customer support feature could be included to

directly process user feedback from real-time usage situations (Oyebode et al., 2020; Radovic et al., 2022). Also, if financially feasible, an option to connect to a human clinician in-app would contribute greatly to the effectiveness of the app (Garrido et al., 2019). Next, a focus on user enjoyment, by including features that involve interactive content, gamification elements, and social aspects (such as peer-to-peer counseling, social circles, or forums) could help increase user engagement. This topic is extra important as our end-users might lack motivation to use the app due to their disorder. A high uptake level could also be achieved using AI models focused on engagement levels. However, the app's helpfulness in supporting the user's mental health should be prioritized over making them spend as much time using the system as possible. Further, there should be a focus on implementing additional features to make the app more appropriate for different users. A more personalized experience can be realized by including a functionality to create and manage a personal health care plan, as shown in the "dream application" in part one. However, it would be useful to have standard versions of healthcare plans available for those who find it difficult to decide on a trajectory themselves. Additionally, an in-app system to provide feedback on the personalized content to improve the app's recommender model could further enhance appropriateness. Lastly, functionalities that enhance a user's trust in the product must be included. It is recommended to add user-autonomy enhancing features, leading to the user feeling in control, rather than being reliant on the system. Also, a noticeable level of human oversight should be included, enhancing the user's trust in the provider and the safety of the system. Other recommendations for obtaining trust will be discussed below.

Expert & stakeholder involvement

Before implementing these features, the first step in the development process is to obtain a group of experts and other stakeholders. They should be included in the whole process, from research to re-evaluation (Schueller et al., 2019; Wykes et al., 2019). Different stakeholders fulfill different roles during the creation of the app. For instance, the inclusion of UX designers is needed to ensure the ease of use of the app, focused on navigation and appealingness (Prakash & Das, 2020). Further, mental health experts and psychologists are needed for oversight of the product's safety and the user's well-being (Titov et al., 2019). They should provide high-quality content, that is relatable, relevant, appropriate, stigma-free, varied, and credible to increase the application's performance and usability. All assumptions and content in the app should be based on existing research and up to standards with acknowledged psychological material. Also, app developers and AI experts are needed to implement specific functionalities and ensure a technically robust and stable product, that meets performance expectations. Next, legal experts should ensure various guidelines and regulations are met during the creation process. There should be a focus on privacy

and consent issues, as well as a safe, ethical, and legal implementation process (Lovejoy, 2019; Wykes et al., 2019). Lastly, end-users must be involved in every part of the process, for the product to fit their needs and wants and help close the gap between research and real-life usage. This could be done by shifting the research focus to clinical research right from the start and doing all testing in real-life settings.

Safety, transparency, & explainability

The next set of recommendations concerns the topic of *safety* for the end-users' well-being and human rights, relevant to many requirements. For example, creating a safer product not only leads to user safety, it also results in a more trustworthy product, that is legally sound and performs better regarding the end-users' needs. The first recommendation made regarding safety is the inclusion of sufficient human oversight (Lovejoy, 2019; Wykes et al., 2019). Humans should be *in* and *on* the loop to provide clinical governance and monitor the AI models. Specifically, the quality of the model's outcomes, such as its insights and content recommendations, needs to be tested and monitored (Wykes et al., 2019). Further, psychologists need to stay involved to monitor user well-being, based on the user assessments. This is especially important considering the sensitivity of the end-users and their receptiveness to harm. Next, the product needs to satisfy relevant regulations and directives, such as the GDPR - to diminish privacy risks - the AI Act's steps for high-risk AI during the application's development, and other guidelines to ensure a responsible implementation of AI. By doing this, some fundamental human rights can be protected and safety issues users previously had with other digital mental health interventions can be resolved.

Related are recommendations concerning *transparency* and *explainability*. Similar to the safety recommendations, a high level of transparency will lead to more trust. Moreover, it forces the development team to deliver a high-quality product and satisfy other requirements such as creating a scientifically valid and technically robust product. To obtain this, maintaining a high level of transparency and explainability (i.e. informative, clear, and concise) is recommended on the following themes: (1) Data storage and processing (Hudson et al., 2022), (2) informed consent matters (Wykes et al., 2019), (3) the working of the model and its outcomes, (4) measures that are in place to ensure safety (Prakash & Das, 2020), and (5) the validity of the provided content (where do claims come from, etc.) (Vo et al., 2019). Related, the product owners should actively take responsibility, as well as be explicit about their level of responsibility for the outcomes of the app (Wykes et al., 2019). Showing a level of accountability taken for the product's consequences again increases safety and the user's level of trust.

Technical feasibility & robustness, and accessibility

Two more requirements deserve more attention in this recommendation section. Firstly, the topic of *technical feasibility and robustness* is an important factor for the system's success. The most important recommendation concerns keeping the implementation plans and expectations realistic: Do we have the ability, technical capacity, and data to create the high-level models necessary for the level of personalization we are after? And if not, how can we adapt these ideas to implement something functionally close to what we want to achieve? Also, what measures do we put in place to ensure the robustness of the solution? These questions as well as additional recommendations about model design and obtaining datasets will be discussed in more depth in the implementation plan below.

Further, it is recommended to perform an environmental assessment before implementing the AI model(s). By lowering the product's needed computational power and emissions, more users can be hosted for longer, making the app more sustainable and accessible. In some countries completing such an assessment is already obliged³.

The last recommendation provided regards the topic of *accessibility*. To overcome common accessibility issues, it is recommended to create and maintain the app using state or EU funding, so using the app can be without pay. Further, it is recommended to incorporate the app into either an existing popular platform (e.g. Instagram) (Schueller et al., 2019) or an operating system such as iOS or Android as an application, increasing the app's exposure and ease of access. Lastly, the creation of a web-based version of the application could lead to more ease of use and uptake, depending on user preferences.

To conclude, these recommendations show us which relevant primary and secondary functionalities can be implemented to create a product effective in supporting students in managing their mental health. The product will be created in a technically robust and safe manner, fueled by the insights of relevant experts and stakeholders. Resulting in an appropriate, safe, and accessible product for the end-users.

Theme	Recommendations	Contributes to requirement:
Primary functionalities	Include the following features: logging/tracking, educational content & exercises, settings to personalize, reminders/notifications, a chatbot, and a crisis feature.	(1) App performance & Usability, (2) Engagement, (3) Personalizability

³ For example in the Netherlands, as recorded [here](#).

Secondary functionalities	Include the following features: In-app customer service, option for clinician support, interactive content, gamification elements, social aspects, management of personal health care plan, feedback option, autonomy-enhancing features, and a noticeable level of human oversight.	(1) App performance & Usability, (2) Engagement, (3) Personalizability, (4) Obtaining trustworthiness, (5) Technical feasibility & Robustness, (7) Closing the gap between research and in-practice usage
Experts & stakeholders	Include (at least) the following experts and stakeholders in the whole process, from research to re-evaluation: UX designers, MH-experts/psychologists, app developers & AI experts, legal experts, end-users.	(1) App performance & Usability, (4) Obtaining trustworthiness, (5) Technical Feasibility & Robustness, (6) Creating a scientifically supported product, (7) Closing the gap between research settings & in-practice usage, (9) Legislative requirements
Safety	(1) Include sufficient human oversight over the model's outcomes and user well-being. (2) Make the product satisfy relevant regulations/directives.	(1) App performance & usability, (4) Obtaining trustworthiness, (5) Technical feasibility & Robustness, (6) Creating a scientifically supported product, (9) Legislative requirements
Transparency & explainability	(1) Maintain a high level of transparency and explainability on the following themes: data storage and processing, informed consent, the model's workings and outcomes, safety measures, and content validity. (2) Be explicit about the product owner's accountability.	(1) App performance & usability, (4) Obtaining trustworthiness, (5) Technical feasibility & Robustness, (6) Creating a scientifically valid product, (9) Legislative requirements
Technical feasibility & robustness	(1) Keep implementation plans and expectations realistic and adapt ideas accordingly. (2) Put measures in place to ensure robustness. (3) Perform an environmental assessment.	(1) App performance & usability, (4) Obtaining trustworthiness, (5) Technical feasibility & Robustness, (9) Legislative requirements
Accessibility	(1) Acquire EU or state funding. (2) Incorporate the app into an existing platform or operating system. (3) Create a web-based version of the application.	(1) App performance & usability, (2) Engagement, (4) Obtaining trustworthiness, (8) Accessibility

Table 2. Overview of main recommendations per theme and the most important requirements that they help meet.

3. Implementation plan: creating a core version

In this implementation plan, we propose a project-like, concrete plan to start implementing AI to support students in managing and improving their mental health. The plan combines AI's

possibilities with our patient-forward studies from part one and our requirements and recommendations for effective implementation. By doing so, we provide a stepping-stone to the development of a core version of an application that includes the primary functionalities wished for by the end-users.

Creating a product like this is a process that develops and improves over iterations of versions. This plan presents the ideas to achieve a minimum viable product. After, there can be built upon the product to satisfy the complete set of requirements and include all necessary functionalities. Through testing and reviewing with the help of our end-users and other stakeholders, we can adapt, add, and take away features.

Earlier, it has been shown that AI brings many opportunities that could be implemented to support the end-users. However, these opportunities can come with risks to the users' health and well-being or are ethically ambiguous. This plan will only use methods that contribute to creating a functional product embodying the core of our aim, responsibly and safely, keeping the end-users' best interests in mind.

The implementation plan will have the following setup: It will start with quickly recapping some recommendations relevant to the first implementation. After, plans are presented to implement six primary functionalities. Each of these functionalities comes with an input-output specification and suggested models/AI techniques that can be used by the programming team to start the implementation. We will also show each functionality's relevance and worth to the overall goal by exemplifying how they contribute to supporting the users. This will be further enhanced by connecting back to our earlier case scenario describing Ana's MH issues, where we will show how using an app that includes the primary functionalities impacts her situation. Further, we will analyze the plan's technical feasibility, address technical bottlenecks, and discuss what research and development is needed to further enhance the implementation. Lastly, we will set up a functional work plan, that can be followed to overcome technical bottlenecks and implement the primary features. All in all, this section reflects on the opportunities and strengths of AI-based mental health support by proposing a practical plan ready to implement.

Recommendations recap

The most important recommendation to follow is to implement the primary functionalities contributing to the overall usability of the product. But before the AI approaches that help realize these functionalities can be implemented, a team of experts and stakeholders needs to be gathered. Firstly programmers to develop the product and implement the AI models are needed. Next, mental health experts to provide data, content, and oversight. We also need a legal expert to ensure

compliance with regulations during the development process and a UX designer for the front-end of the application. Lastly, end-users should test and provide feedback on the product on multiple occasions. Also relevant are the recommendations on *safety* and *transparency* during this process. This will contribute to responsibly and ethically creating the product. Recommendations on technical feasibility will be included in the feasibility analysis below.

Proposed primary functionalities

To create a minimum viable product of an application effective in supporting students to manage and improve their mental health, six primary functionalities need to be implemented. For each functionality, a quick description is provided, followed by an illustration of the potential output that exemplifies how it functions and why it is a valuable feature of the app's core version. After, a possible AI approach is presented that can be used to realize the functionality's implementation. We describe how the model should function, what techniques could be used, provide input and output specifications, and discuss some potential pitfalls. After this description, you find a table that summarizes the functionalities and their proposed AI approaches (Table 3.).

A. Logging and tracking feature

This feature allows users to track all kinds of data points like symptoms, moods, events, daily habits, and more. The app will provide insights and statistics on the tracked data points, could help reflect on those insights and give suggestions accordingly.

Example The app can, for example, give insight into a correlation that negatively impacts the user's mental health, and follows that by suggesting an activity that correlates with a positive result: *"I have noticed that when you spend less time on TikTok, your mood is often lighter. Would it be a good idea to go for a walk instead of continuing to scroll, doing so often diminishes your feelings of anxiety."*

AI approach Two approaches to implementing this feature would be: 1) Either to use an *unsupervised machine learning model*, to find patterns in the tracked data points. This model would use user data as its input, and outputs relations between data points, predictions based on earlier found patterns, and suggestions based on the relations and predictions. A downside to this approach is the fact that an unsupervised model needs a lot of input data to produce helpful outputs. To the amount that a single user cannot deliver. Also, the model would have no notion of desired vs. undesired mental states (e.g. calm vs. anxious). Without this, it would be impossible to give helpful suggestions.

That is why 2) *a semi-supervised machine learning model* would be a better approach. We

recommend creating a hybrid model, that is at its core unsupervised – and finds hidden patterns - but is enriched with a “knowledge model”, containing extra data such as a notion that a happy mood is desirable, known connections between data points, and labeled examples of relations that might indicate a specific disorder. The model’s outputs would be similar, but more guided by what is known about mental disorders.

For the second approach, a pitfall would be that a considerable amount of data is still needed. Ways need to be found to expand the dataset without losing quality. There is also a risk of (confirmation) *bias* with this second approach. Who decides what patterns are desirable? Will we miss important connections because we only focus on known relations? Though important to remain aware of these weaknesses, the combination of domain knowledge and unsupervised ML would be the recommended approach, accompanied by an investigation of missed connections or potential bias.

B. Educational content and exercise recommendations

Throughout usage, the app will recommend educational content, psychological/therapeutic exercises, and meditation sessions. Further, it will guide and support the user through the exercise or content. The content suggested is appropriate to the user’s symptoms, clinical picture, and other needs (e.g. normalizing MDs, or building confidence).

Example The app might have noticed that a user has developed strong feelings of self-criticism towards coping with their mental disorder and in life. To counter those feelings, the app suggests practicing self-compassion and presents some exercises that might interest the user as they also fit their ethnic background.

AI approach The AI approach that can be used to implement this feature is a *supervised, machine learning-based recommender system*, where certain data points and relations lead to certain content recommendations. Additionally, users could be clustered to make training the models less computationally demanding and have more input data available to train them with (from a user cluster rather than individually), giving us more robust outputs. Doing so prevents us from spending many resources on single users, rather than creating something beneficial for many users (Vlek et al., 2012).

Data on relations found in the logging and tracking functionality, combined with additional user input from self-assessments, and possibly even in-app behavior, help us to categorize users. The users get labeled with information like the type of disorder, preferred therapeutic strategy, additional needs (e.g. confidence or motivation), and relevant personal information (e.g. ethnic

background or sexuality). Educational content and exercise sessions will be recommended based on those categorizations.

Though creating a user profile is promising, we must be aware not to create “recommendations bubbles”. We face the risk of putting the user in a box; giving them less room to explore, learn, and give input into their care plan. To make sure this feature remains supportive and user autonomy is maintained, the system should always present diverse suggestions and give the user room to make their own decisions.

C. Reminders and notifications

The user will receive notifications to help them stay engaged and remind them to work on their mental health. The notifications will be personalized and depend on the user’s care plan, personal preference about the intensity and content of the notifications, and linguistic style.

Example Notifications from the app could be: *“I noticed you often log ‘a lack of social contact’ and ‘somberness’ together. Want to come over to explore that correlation?”* or *“Time for your weekly journaling session”*.

AI approach The proposed model for this functions as a classification model, that provides textual output based on the tracking and recommender functions, making the output personalized. Similar to the content recommendations, the reminders and notifications are also dependent on the user categorization done before and take into account information on the user’s background, clinical picture, and symptom tracking. The user’s in-app progress can also be included to determine relevant information to notify about. For example, if a user chooses to participate in a specific series of exercises, the reminder should be about the right exercise. Further, the user’s personal preference about the amount of notifications and their content (as indicated in their personal care plan), should be part of the input data for the production of notifications.

Additionally, the produced texts can be tailored to the linguistic style of the user and mimic the user’s own use of language, for a higher appeal.

It should be noted that if there are concerns about computational power, the notification feature is a good place to save resources. Not all notifications need to be dependent on statistical AI methods such as machine learning. Implementing, for example, a lighter rule-based model could be enough. This also prevents over-tracking and an invasive (big brother) effect. If the notifications functionality is implemented using ML after all, it would be ethical to regularly remind the user of what data is used to notify and recommend content and ask whether they would like to take out information stored in their user profile.

D. Chatbot

The chatbot - integrated into several features in the app - has varying functionality. It can function as a reflective tool, to help reflect on the tracking functionalities, how the exercises and learnings are going, or simply on the user's current mental state. Simultaneously, it can function as a decision-aid, to help users plan and get through their day if making decisions is difficult, or simply for them to have a compassionate, empathetic conversation.

Example A user just participated in a journaling exercise.

Chatbot: *"I noticed that you didn't write as much as you normally do. Is there a reason for that?"*

User: *"Maybe.. I guess I was distracted because I had a lot going on today."*

Chatbot: *"I'm sorry to hear that, anything you want to vent about?"*

User: *"I took my anger out on my girlfriend, and I regret it".*

Chatbot: *"Hmm regret is a difficult feeling. How did you act afterward?"*

The chatbot allows the user to talk about their internal state, which could help them clear their mind and possibly reflect on what happened today, how they felt about it, and how they want to approach similar situations in the future.

AI approach In different features in the app, an intelligent chatbot that can have varying conversations is required. Sometimes a user wants help reflecting on their internal state, thoughts, or events in their life, sometimes they simply seek a compassionate voice or someone to listen to them.

To build a sophisticated chatbot like this, a solid foundation is vital. Therefore, we suggest using a strong Large Language Model (LLM), such as GPT-4o or Mistral 7B, as the foundation model, and potentially finetuning it for this specific purpose ([OpenAI, 2024](#)). Rather than tuning it ourselves, it is recommended to leverage other models that are already finetuned to the domain of mental health⁴. From there, the model could be provided with a "context window" consisting of transcribed conversations between patients and therapists, to "teach" the model what kind of responses are expected. Other contextual input data for few-shot learning can be provided by acknowledged psychological sources, such as the DSM-V, to provide information on different disorders, their symptoms, and contexts in which they occur.

With this functionality, it is extra important to be aware of the sensitivity of the end-users. High-level monitoring of the chatbot's outputs is required to prevent harm. The model should be audited carefully by MH experts during training as well as periodical re-evaluation.

⁴ Other relatively strong models can be found at for example huggingface, and a quick search shows some already pre-trained for mental health: e.g., [mental-health-mistral-7b-instructv0.2-finetuned-V2](#).

It must also be mentioned that using GPT-4o would come with privacy concerns, as it is not open source and it cannot be confirmed that user data will not be sold to third parties. Therefore, using a more ethical alternative to OpenAI can be preferable. Other concerns of using GPT include accountability, safety, and bias problems, as well as environmental problems (Stahl & Eke, 2024). With the current state of the art, these are problems that occur with many LLMs.

E. **Personalization options**

The app should have a function to plan and customize one's own care in the form of settings, making the app more appropriate for the user. There could be settings regarding different disorders, different types of content, and different intensity preferences in using the app (such as the amount of notifications or assessment moments). This feature can be extended through iterations of the app, with more adaptable options, and more variation within the options. Also, in this functionality the user will decide what data will be used for the tracking model.

Example A user suffering from major depressive disorder gets different content recommended by the app than someone with an anxiety disorder. One might be more receptive to cognitive behavioral therapy, whereas the other gets better results from acceptance and commitment therapy. One prefers to interact with the app multiple moments a day, whereas the other only wants to work on their mental health in the mornings.

AI approach In the first iteration of this functionality, an AI model does not have to be involved. This feature functions as a "settings" functionality, to indicate a user's wishes about their app usage and care. The information in this feature will function as input data for the app's AI model(s) used to recommend content or make its chatbot more appropriate.

In later versions, it is desired that a user can fully adapt and manage their care plan using this feature. For a later, more sophisticated, version it is suggested to use the AI model containing the user profiles, to help the user adapt their care plan with insights from other functionalities, such as logging, progress with the exercises, or even findings from conversations with the chatbot.

More personalization means higher computational costs. It will be a challenge to create a more appropriately functioning product while controlling the app's sustainability.

F. **Crisis functionality**

This feature allows both the user and the system to intervene at a user's low point. Either the user or the system can detect and indicate when there is an emergency and - depending on the user's preference at that moment - help is offered. This could be in the form of being connected to

a peer (to chat to) or to be guided through a specific exercise relevant to the crisis. The system learns from these crisis moments and will intervene when a crisis might be present or lie ahead.

Example A user suffering from depression is using the app to practice positive thinking. The model notices the user is being less engaged than usual and is writing down some concerning words, related to an earlier heavy wave of depression. The app asks the user if they are okay with stopping the exercise and answering some questions. By doing this, the app tries to determine whether the predictions are correct and what the best course of action would be to help the user now.

AI approach There are a few ways to realize a crisis feature. The first method is to put more functionality into the statistical tracking model described before and give it the additional task of *predicting crises*, based on prior user indications of crises and connected data points.

Apart from statistical machine learning methods, techniques from Natural Language Processing (NLP) - such as named entity recognition (NER) or comparable methods - could be used to perform word detection. Based on psychological research a list of “heavy words” could be composed, that represent patients’ low points, and have the system match these throughout the entire app. From there, a classifier or decision tree could be used leading the user to their preferred support option. This solution is less computationally heavy compared to statistical machine learning, and could therefore be a nice alternative to achieve the same result more sustainably.

There are a few challenges that come with these methods. Firstly, when using statistical machine learning, much data is required. Learning from only a small set of prior user indications of crises can be challenging. More data will need to be gathered that will improve the accuracy of crisis predictions. A second problem, relevant to both the ML model as well as word detection is the occurrence of false positives; crises detected by the system that are not actually crises for the user. They can be perceived as annoying and result in disengagement. A simple solution would be to include a feedback system, that updates the system’s knowledge model on crises based on user feedback.

Combining the models

As a final remark, it should be clear that models proposed for the separate functionalities could be mixed or combined. The tracking model and recommender system from functionalities A and B could be woven into a single model, while also being integrated within the chatbot. Similar to how the crisis functionality can be a part of the tracking model, combined with NLP techniques. In the final product, approaches could be combined and mixed for feasibility purposes and to create an end product that all all-encompassing.

Feature	Function	AI approach and (technical) challenges
Logging and tracking	Tracking of various data points, providing insights and statistics, and making suggestions accordingly.	<p>AI Hybrid, semi-supervised machine learning (ML) model enriched with knowledge model to find patterns and connect to psychological knowledge.</p> <p><i>Input:</i> User data, knowledge model.</p> <p><i>Output:</i> Relations between data points, predictions, and suggestions presented narratively.</p> <p>Challenges Bias, generalization, limited data available.</p>
Educational content and exercise recommendations	Recommendation and guidance of educational content and therapeutic exercises.	<p>AI Supervised ML-based recommender system that matches relations in data to content recommendations for different user clusters.</p> <p><i>Input:</i> Patterns from tracking features, additional user input from self-assessments, and app usage.</p> <p><i>Output:</i> Content recommendations presented narratively.</p> <p>Challenges Recommender bubbles, high computational costs, loss of user-autonomy.</p>
Reminders and notifications	Help the user stay engaged by providing relevant, personalized notifications.	<p>AI Classification function that provides recommendations based on tracking and recommender model.</p> <p><i>Input:</i> Suggestions from the recommender model, known user information, and app-usage data.</p> <p><i>Output:</i> Textual, personalized reminders/notifications.</p> <p>Challenges Computational costs, over-tracking.</p>
Chatbot	Chatbot with varying functionality: such as to help reflect or make conversation.	<p>AI Finetune a strong LLM as the foundation model to the domains of MHC and therapy.</p> <p><i>Input:</i> Foundation model, transcribed conversations of therapy sessions, acknowledged psychological sources, user input during run time.</p> <p><i>Output:</i> Relevant and appropriate textual responses to user input.</p> <p>Challenges The ethicality of the foundation model, user sensitivity (risk of bringing harm), and proper auditing.</p>
Personalization options	Letting user plan and customize their own care plan using settings.	<p>AI In-app settings option, that provides input for the app's AI models. The settings themselves do not rely on AI techniques.</p> <p>Challenges Increasing the amount of personalization while keeping computational costs low.</p>
Crisis-functionality	Intervention at the user's low point, when direct help is needed.	<p>AI Either use the tracking model from functionality A to predict crises based on a combination of tracked data points and user indications of earlier crises. Alternatively, perform word keyword</p>

		<p>detection throughout app usage with a list of words representing a user's low point.</p> <p><i>Input:</i> User indication of crises, tracked data points, in-app behavior.</p> <p><i>Output:</i> The detection of a crisis that activates a support plan.</p> <p>Challenges False positive interventions, limited data for proper prediction.</p>
--	--	---

Table 3. Summary of the primary functionalities, the suggested AI approach to realize them (including input and output specifications), and potential challenges.

Ana again

To fictively illustrate how the core functionalities interact with our end-users in their progress to mentally feel better, we present another case example, where we cycle back to Ana. The story shows how the minimum viable product helps meet the users' wishes and needs, and what parts would need to be further enhanced or added next to create an app highly successful in supporting students. The case also shows an example of how the primary functionalities are combined within the workings of the app when an exercise is suggested to our fictive user.

Recall our example case student Ana: she suffered from major depressive disorder with typical symptoms, such as losing interest in her daily activities and feeling sad. Further, she experienced additional feelings of worthlessness and shame, due to her cultural background. As a result, she failed her classes and lost genuine contact with friends and family. When trying out mental health apps, she was mainly frustrated that the contents were not appropriate to her problems and background, lacked the motivation to keep engaged, and sometimes even felt worse after doing the suggested exercises.

Hence, Ana responded to an advertisement from the European Union to become a beta-tester for a new app they are designing to support the mental health of university students.

The current version of the app that Ana has been using for a month now has a few features. She is especially impressed by the *tracking functionality*. It has explained many things about Ana's well-being, as she now understands connections between certain events. The content provided in the app is of high quality; it is provided by experts and uses acknowledged therapeutic techniques. The exercises recommended to her do not scare her anymore, as they are actually helpful. Through interactively working with the content, her views on mental disorders start to change slowly. Also, she acquired some helpful tools and coping strategies that help her deal with her thoughts and feelings throughout the day, giving her a greater sense of control over her mental state. The techniques from cognitive behavioral therapy have helped her go to social events again. Further,

the app made her find the motivation and strength to return to class, by repeatedly following the exercise of bending her thoughts to a more positive outlook on her situation and state.

Most of the content matches her clinical needs, however, she still wishes the recommendations to be more diverse and also match her emotional and cultural background better. This is a piece of feedback she gives back to the development team. The personalization is mainly done through subjective assessment and the settings function, but this could be more extensive and inclusive still. For now, she can decide which symptoms to track, how many exercises and check-in moments she wants and when, and how many notifications. Ana likes the notifications, as less proactivity is required. They are also presented very playfully and relevant to her progress in the app, rather than being generic. This has made her more tempted to click them. To provide an example:

“Ana found herself lying in bed in the middle of the day for a while when the app suddenly popped up with a notification. It read: “It’s okay to stay at home when you feel like you have to! But a little challenge I have in mind might help you spin this afternoon around...”. The app could know her location because she consented to GPS tracking and provided the coordinates for a few different locations, such as her house and the campus. The notification about a challenge made her curious enough to click it. The app guided her through an exercise where she had to perform little actions -one step at a time - and reflect on what she thought and felt while doing them in between. The exercise resulted in her wearing her shoes and coat, setting the bar lower to go for a little walk that was suggested next, or repeating the steps backward – consciously again - if walking outside was a bridge too far. Now that she was already up and dressed, she decided that a little walk to the bakery would not hurt.”

The exercise was led by the in-app chatbot, which is used to show insights on tracking events, recommends content, and offers a place to casually chat. The bot understands Ana’s input and produces appropriate and useful responses in a nice tone. Lastly, the app contains an intervention function that has popped up twice throughout using the app. For some reason, it felt the need to intervene during an exercise and change the course of the exercise to tackle a detected crisis. Ana didn’t really notice yet how down and stuck she was feeling, but the crisis plan made her find to strength to take action and turn it around. Ana felt empowered when she noticed she was taking better care of herself and when she dared to be out in public and have a laugh with friends. Yet, she felt it even more when – through conversation with the chatbot during a detected crisis – she found the guts to pick up the phone and call a friend for help. Though it was scary, it was also very relieving. Ana and her friend now chat every couple of days about how she is doing. The app supported her in re-finding her human support system.

In the past, Ana struggled with staying engaged with MH apps. Despite still missing features that could improve this further, she noticed being more engaged because of the high-quality content provided, the sophistication of the tracking model, and the notifications.

After having used the app for a while, Ana’s feelings of shame started to recede and she decided to become more actively involved in improving the app, joining several meetings to provide end-user feedback on new features. These plans included functionalities like in-app feedback on the recommendations to strengthen the AI model, gamification, and goal-setting features, as well as the inclusion of a social aspect where users get matched into social circles with other testers with similar problems and backgrounds. Ana emphasizes the importance of more personalization options, to enhance the quality of the recommendations. Different users have different cultural or social backgrounds, or different healthcare goals, resulting in different content needs. She also suggests that a basic version of the app with fewer features should be optional, as starting to use the app can feel very overwhelming at first.

This case scenario illustrated how the primary functionalities presented in the implementation plan contribute to supporting students in improving their mental health. As you recall, potential end-users have expressed a set of wants and needs for MH systems. The requirements introduced at the start of part two are designed to cover all user needs and wants. The primary functionalities are based on a single recommendation theme and help to realize a subset of needs/wants. The remaining needs will be realized when all requirements are met.

Table 4 presented below, first shows how a set of user wants and needs is achieved by implementing different primary functionalities (and potentially strengthened by other requirements). Further, it shows how the remaining wants will be featured in later iterations of the product when all requirements are met. The table shows how our findings from part one relate to the recommendations and proposed implementation from part two.

User wants	Primary functionalities	Requirements**
1. Logging/tracking	A. Logging/Tracking	App performance and usability & Personalizability
2. Educational materials and exercises	B. Educational content and exercises	Engagement – primary functionalities
3. Personalization options	E. Personalization options	Personalizability
4. Rewards, goal-setting & gamification*	-	Engagement - secondary functionalities
5. Social connectedness*	-	Engagement – secondary functionalities
6. Human support*	-	App performance and usability – secondary functionalities
7. Fitting into daily routine	E. Personalization options & C. Reminders and notifications	Close gap research and daily lives – secondary functionalities & stakeholder inclusion

8. Enjoyability*	-	Engagement – secondary functionalities
9. Reminders & notifications	C. Reminders and notifications	Engagement – primary functionalities
10. Crisis intervention	E. Crisis functionality	App performance and usability
11. High-quality content	B. Educational content and exercises	Scientifically supported product
12. Anthropomorphized system	D. Chatbot	Engagement – primary functionalities
13. Trustworthy system	-	Obtaining trustworthiness & Legislative requirements
14. Accessibility	-	Accessibility
15. App stability/functioning	-	Technical feasibility and robustness
User needs		
1. Patients planning their own care	A. Logging/Tracking & E. Personalization options	App performance and usability & Personalizability
2. Normalizing MH problems	B. Educational content and exercises	App performance and usability
3. Use of acknowledged techniques	B. Educational content and exercises	Scientifically supported product

Table 4. This table connects each user want/need to either a primary functionality or the requirements that cover them *most strongly*. Of course, most requirements interplay and are relevant to more wants/needs.

* These wants can be found in the set of secondary functionalities, presented in the recommendations section.

** Almost all user wants contribute to the requirement of “App performance & Usability”, however, in this table only the strongest relations are highlighted.

- the dash indicates that the user want is not covered by the primary functionalities.

Feasibility analysis - a discussion

Next, we will analyze the feasibility issues that come with this implementation plan. We have discussed what state-of-the-art AI approaches can be used to realize the features of a minimum viable product, and already identified some weaknesses that require further thinking. What other technical bottlenecks do we face? On what fronts do we need additional research and development? This last section will discuss the feasibility of our plans and lay out the remaining challenges that need to be overcome to bring the implementation to a higher level.

Data: gathering & management

This discussion will start where creating an AI-based model generally starts as well: gathering data. The first important question is: Do we have enough data to create the models and is the data of high quality? The quality of your model is the direct result of the quality of your data. There needs to be enough data for proper training, and it needs to be diverse, consistent, complete, representative, and relevant regarding the subject.

In the AI approaches described above, we have discussed the necessary input specifications, such as user input data from live usage, or transcriptions of conversations between therapists and their clients. Sometimes when applying AI techniques, researchers face the problem of data shortage,

and due to the sensitivity of the subject, this will likely also be the case for us. There are a couple of ways to deal with this problem. For example, making use of alternative data streams to create a knowledge framework and add that to the dataset. Mental health experts and practitioners can help us select acknowledged sources, existing exercises, information on therapeutic techniques, diagnostic means, and provide other insights. A second option, to improve training the semi-supervised ML models, is to create synthetic data by either leveraging LLMs (Josifoski et al., 2023) or using data augmentation techniques, such as rephrasing and remixing existing text (Organisciak & Ryan, 2022). The synthetic data will resemble the current data and help increase the size of the dataset, without being a risk to the users' privacy. A downfall is that this approach requires a lot more pre-processing to make sure the synthetic data is of high quality. Yet, the work may be worth it. Lastly, the use of "Differentially Private" (DP) training methods ensures sensitive data remains private when training ML models (Kurakin et al., 2023). By using this technique we could use more sensitive data as privacy can be ensured.

However, there is no need to reinvent the wheel; we could also utilize pre-trained models for similar domains and fine-tune those with the data we do have. This method requires a lot less data. Lastly, sometimes we are forced to adapt the plans to the available data and use other means to create models with similar output. Based on the available data, the AI experts in the team need to research the different options to achieve functionalities helpful to the users' aim.

To responsibly create a product, a high level of data management and security needs to be ensured. Because there is an aim to handle the data ethically and respect the users' privacy, the responsibility that comes with that directly impacts the technical feasibility of our plans. The dataset for this project is highly sensitive and personal and needs proper security. Some methods described above, for example, using GPT as the foundation model or the leveraging of other models might bring even more privacy concerns. However useful these methods and models are, protecting the users' privacy has to remain a top priority. Therefore, there is a need to research the ethicality of the third-party models aimed to be used, and potentially find new solutions if they do not meet ethical standards.

Computational costs

For the sake of this paper, the assumption can be made that the team gathered has the ability and skills to create the models described above. However, an assessment is needed to determine whether the technical capacity to create and maintain these models is available. Building, training, and continuously running machine learning-based models or neural networks can be a costly process. Computational power is needed to run the models on servers, as well as to store and secure the increasing amounts of user data. New data provided through app usage will

have to be processed by the model, using large amounts of computational power. Not only is this a monetary costly process, but it will also produce lots of CO₂, making it costly for our planet. As recommended, an environmental impact assessment is required. From there we need to balance how much computational power we want to use up against what good the model will bring for our society.

There likely is a need for ideas to reduce the environmental footprint and computational costs of the product, as this will increase the project's feasibility and make it more sustainable to use in the long run. The development team needs to research solutions that will contribute to the reduction of computational power. Ideas that can be researched are, for example, the clustering of users (as mentioned in functionality 2) to only adapt the model based on data gathered from an entire cluster. This will make the recommendations and insights less personal but will save us computational costs and make it possible to run the model for large amounts of users. Also, as mentioned before it is not necessary to use heavy statistical machine-learning solutions for every functionality. Implementing the crisis functionality or reminders and notifications can also be done using other types of AI techniques that require less computational power.

However, it is required to add even more personalization in later versions compared to the proposed core version, as that is one way in which AI-based solutions make a difference compared to current DMHIs. The question remains how we will achieve such a high level of personalization, whilst keeping to an acceptable level of resource usage. There is a need for further research to adapt the plans and realize the selected functionalities in an eco-, and cost-friendly manner, to be able to help as many students for as long as possible.

Robustness

Further, after tackling the feasibility issues discussed above, we find a need to ensure the product's robustness. Creating a robust product is highly important for the user experience, but also prevents mistakes that can bring harm to sensitive students. Making robustness a valuable factor for safety. Also, a robust product is more trustworthy and therefore contributes to the overall effectiveness. App developers and AI experts need to add measures that enhance the product's robustness.

As mentioned briefly while introducing the AI approaches, there are a few more technical risks and ethical weaknesses to consider. We also need to overcome pitfalls like bias, and recommender system bubbles, as well as prevent over-tracking and invading the users' mental state. All in all, more technical research to implement the patient-forward functionalities ethically and responsibly is required. Before further technological advancements, plans need to be adapted

to the current state-of-the-art, and extra measures need to be undertaken to ensure a feasible and robust product that is helpful to the users.

Work plan: step-by-step implementation suggestions

Up until now, we have presented a set of requirements that need to be met for AI-based MH support systems to be successful in helping their end-users. After, themes of recommendations have been presented to help realize them. One recommendation theme concerned proposed primary functionalities that combine into a *minimum viable product*. In the “Ana again” section we illustrated how these functionalities can support students in improving their mental health. Contrary to the sci-fi-esque dream scenario, this implementation suggestion finds its foundation in existing, state-of-the-art techniques, making it realistic to create an app that includes the primary functionalities to help students like Ana.

In this final section, we will propose a concrete, practical work plan that can support a development team to start the implementation of the proposed primary functionalities. The work plan includes more concrete AI suggestions presented in a step-by-step approach, taking into account the remaining feasibility issues, to ensure a realistic project process and realize a functional product. For this part of the paper, we assume that a team of experts and stakeholders has been gathered to play their part and add their expertise to every step of the process. The work plan will mainly focus on the AI experts’ and developers’ roles. We will, for example, presume that a legal expert ensures compliance with the safety steps of relevant regulations, such as proper data management. Also, note that the work plan steps are created to provide guidance and include mainly suggestions. It is recommended to assess whether the use of different models, different features, and additional steps is necessary.

Recall that it was suggested before to create a combined model that encompasses all proposed AI-based functionalities, to create a combined output presented in “Ana again”: The need for the *suggested exercise* was discovered through the *tracking functionality*, for which Ana *personalized* to track GPS data. The exercise was suggested to her via a *notification* and carried out by the *chatbot*. To achieve output like this, the used AI techniques need to be either combined within one model or be used in interlinked models.

Work plan steps

- 1. Create personalization options:** As most of our proposed methods are driven by user input data, creating a way for the user to provide that input will be the starting point of this project. The personalization feature consists of an assessment of relevant background information supported by questionnaires from the MH domain, wants from the app in

delivering support (e.g. the number of exercises or check-in moments a day and when), and what kind of data points to track (recall various options in Table 1.). It is important to provide a fixed list of options for the user to track, rather than having each user type in their answer. Some of these data points need to be filled out manually by the user (e.g. “today’s mood”), while others are tracked continuously (e.g. “social media usage”). Doing so will create a first *user profile*, consisting of numerical “scores”, that functions as a knowledge model to input into other functionalities. Remember to prevent over-tracking and provide the users with more autonomy by regularly reminding them what is being tracked and asking if they would like to make any changes.

- 2. Data gathering and pre-processing:** To create and train the semi-supervised ML model used for (1) providing insights into personal user data, as well as for (2) recommending content and later to potentially train (3) the chatbot and (4) the crisis-functionality for more personalized service, enough, high-quality data needs to be gathered. The field of psychology can provide datasets on known symptom relations, patterns, and personality/disorder assessments. This data needs to be diverse and representative to decrease the risk of bias. Also during pre-processing and categorizing, data points that can cause bias need to be removed. Depending on how limited the data collection will be, synthetic data will need to be created, making the dataset more substantial. Alternatively, a fitting model pre-trained for a similar domain needs to be found in step 4, so that it can be fine-tuned with the limited dataset.

For the tracking model, we are using a fixed list of moods, events, and symptoms including data points such as: “feeling hopeless”, “tired”, and “did not leave the house today”. By doing so the user input can be converted to numerical representations - e.g. using one-hot encoding - to process the data more effectively and help identify patterns and relationships between the data points. Using one-hot encoding, every symptom could be weighted equally, and easily expanded by adding labels. Training and running the model with numerical data also lowers the computational complexity and enhances the model’s accuracy (if the semantic meaning is captured properly). The dataset needs to be split into training, validation, and test sets. To create a semi-supervised model, it is needed to give the data some additional labels indicating symptom severity (positive or negative, or even assign scores). Also, pre-defined relations between data points and classifications of what those relations mean need to be added, to give our model more guidance.

Lastly, to add the recommender functionality, content must be labeled with their connection to certain symptoms, events, relevant background, patterns, or relations.

3. Feature engineering, model selection, and training: The first functionality that needs to be implemented is the “logging and tracking” functionality, as it functions as a base for other features, such as recommending content, sending notifications, creating a base for the chatbot’s output, and potentially the crisis functionality (in a later version).

To help the model along in finding patterns and relations, relevant features need to be designed to enhance training. Examples of features that could help discover relations between logged data points could be: occurrences of specific symptoms (or sets of symptoms) or time-related patterns (i.e. when the symptoms occur (compared to other symptoms)).

Different types of models could be selected to implement this functionality, depending on the eventual size of the dataset and computational resources available. To achieve outputs on correlations between data points, a correlation matrix might already be sufficient.

Further, the user profiles could function as input to a simple neural network, to discover hidden patterns and more complex insights. To include the temporal aspect of tracking symptom increases/changes over time, a Recurrent Neural Network (RNN) could be a good option. More specifically, Long Short Term Memory (LSTM) can help us keep track of important data points over time while limiting the amount of data storage needed (Ramu et al., 2023). Further, we could choose to use clustering techniques to cluster data points together and score users on what clusters – or new relations - can be found. Later, these clusters can connect to content recommendations. When choosing a model, it is suggested to aim for a balance of using existing knowledge while allowing the discovery of new insights and relations from the data. The model you choose should reflect this aim.

An important step in the training process is to regularly evaluate the model and update it with new data to discover evolving patterns. There should also be active monitoring for biased predictions, using metrics that assess fairness and accuracy or during real-life testing to make necessary adjustments. Further, it is recommended to choose a model that is less prone to overfitting and use bias correction techniques, to mitigate the risk of bias.

Lastly, if the choice is made to use a pre-trained, third-party model, make sure that it meets ethical standards and respects and protects the users’ privacy.

4. Environmental impact assessment: As training and running the model(s) that realize a high level of personalization demands lots of computational power, performing an environment impact assessment is strongly recommended. The solution’s needed amount of computational resources needs to be balanced against the usefulness towards our goal. Potentially, we can think of ways in which we can adapt the plans to reduce as much computational power as possible, without losing quality. Solutions that will contribute to the

reduction of computational power include: the use of pre-trained models, user-profile clustering, reduce of statistical ML solutions when other, less demanding, models can be used instead.

5. Design outputs and combine the tracking model with other functionalities: Next, we need to design outputs. This means relations between symptoms, patterns found, and other insights need to be translated into textual messages, understandable and relevant to the user. Further, the model's outputs (relations, updated user profiles, etc.) need to be incorporated into the user profile and therefore used as input for the other functionalities. These will be discussed briefly below. Generally, all model outputs should be presented textually.

a. Content recommendations: Based on the data point relations and patterns found, the user profiles will be updated. Additionally, the user's in-app progress - for example, information on articles that have been discussed or exercises that have been done – must be added to the user profile. These profiles indicate which data will be recommended to the user. By doing so, the recommended content will stay relevant and repetition will be limited. Achieving this is possible by using a rule-based model or decision tree, where certain weights or combinations of tags present in the user profile, lead to different recommendations.

To reduce computational costs, users can be sorted into user groups. Content recommendations will then only need to be calculated for an entire group. Further, to avoid creating recommender bubbles, diversity methods can be introduced; related content that does not fully match the user profile can be presented to breach out of the bubbles. Lastly, a feedback option can be included that adapts the user profile, to give the user more control over the content recommendations.

b. Notifications: As a first step, the notification functionality can be achieved with a similar or even connected rule-based model, that notifies the user based on information present in the user profile. The notifications can be about insights from the tracking model, content recommendations, or even general engaging messages. For this feature, the outputs created thus far will be translated into a style fitted for the role of a notification or reminder. Large language models (LLM) can help us translate that output to create variation.

c. Chatbot: As a first step, an in-app bot can be introduced that provides one-way communication on findings and content, with no option to respond or react. This creates a platform or way to present the outputs from the tracking model and connected content recommendations. The textual outputs should already be presented in an anthropomorphized, chatbot-like manner. In the next iteration, the user could be

given two or three fixed options to respond to an output, modeling the chat function with a decision tree. Different user-response options lead to different conversations and different presentations of tracking outcomes and recommendations. After that, the chatbot can be built and tuned on a LLM foundation model making it able to uphold a proper conversation with a user.

The chatbot's messages should either be monitored by mental health experts, to ensure user well-being. Alternatively, to enhance the chatbot's quality "Reinforcement Learning from Human Feedback" (RLHF) can be introduced. This is a method that can strengthen an NLP model by optimizing its outputs based on human feedback (Javaid, 2024).

- d. **Crisis-functionality:** For the first version it is recommended to use a separate model that responds to user inputs in self-assessments or potential textual input during exercises. Using insights from MH-experts, a list of keywords or phrases can be created that call for intervention, and start a "crisis plan". The crisis plan can be created with a decision tree, giving the user action-options dependent on their needs. The key-word detection can be executed by using a Named Entity Recognition (NER) model, or comparable.

In a later version, the interventions can be connected to the tracking model to create a more personalized service, where some detected relations or events trigger the start of the crisis plan. An example of this functionality was shown in "Ana again", where her GPS data triggered the system to recommend a specific exercise. Interventions can be triggered similarly.

Conclusion

In this thesis, we researched how AI can best support students in improving their mental health before human help is available. From a patient-forward perspective, we analyzed the general desirability of AI-based mental health support systems, the end-users, their needs and wants, current digital mental health interventions, and AI's opportunities regarding this subject. From there, we theorized about a dream application that formed the basis for a practical framework in part two. In this framework, nine requirements needed for the effective employment of AI-based support apps were introduced. Additionally, the requirements were supported by recommendations to help meet them. Further, we tried to close the gap between psychological research into the effectiveness of digital mental health interventions and the actual development of such systems, by directly applying the learnings from the first part into a practical implementation

plan that can be used to create a patient-forward, well thought-out, AI-based MH-support application.

The analysis showed us that there are undesirable ways to employ AI for mental health care, namely when it interferes with or functions as a replacement for human care, or is employed as a mere *technofix*. An MH-support system should not take away the user's autonomy, force, or manipulate them. However, if used as an additional supportive tool - whose aim is not to improve students' mental health directly, but to support *them* in improving their MH *themselves* - there is a place for it. The AI-based system should *empower* users and provide them more control and autonomy over their mental state, whilst not harming their well-being and respecting their privacy. AI's strength lies in improving the appropriateness of DMHIs by increasing the level of personalization. In doing so, it distinguishes itself from other technological applications.

Also, a preference for the use of mobile applications was found, because of their wide accessibility and potential to support users any time of day. Applications are found to be effective if they meet the following *requirements*: they need to be helpful, engaging, personalizable to the user's care needs and background, trustworthy, technically robust, scientifically valid, fit in the user's daily routine, accessible, and lawful. To fulfill the users' needs and wishes, it is recommended that the application includes (at the very minimum) the following functionalities: a logging and tracking feature, educational content and exercise recommendations, reminders and notifications, a chatbot, personalization options, and a crisis functionality. All functionalities aim to enlarge user empowerment and autonomy, by simultaneously guiding them in working on their mental health.

The design requirements shown in this thesis are not merely a wish list. All of these features can be responsibly realized with state-of-the-art AI, as indicated in the implementation plan. However promising, there is still a need for more technical research and development to overcome remaining challenges and create high-quality, sustainable models, that meet all of our requirements. Yet, by additionally introducing a work plan, this thesis provides a stepping-stone for a first implementation.

In this research, a patient-first approach was chosen to help us reason about what is necessary to support the end-users instead of what is technologically possible. In doing so, we hoped to find facilitators and barriers to the effectiveness of existing mental health systems and create a solution that truly matches the users' needs and wishes.

As mentioned before, this research could benefit from adding the perspective of mental health experts, clinicians, and other stakeholders, giving the project more dimension. Also, it should be researched how much "the societal debate" surrounding technologies applied to this sensitive domain influences the effectiveness of our proposed framework. Further, even though the

effect of the users' mental health problems on their ability to engage with new systems has been tried to incorporate, this might be an even bigger factor than assumed in this piece. Future researchers need to remain aware of this factor when testing and improving the proposed product. Lastly, though at times this thesis may seem to be an extensive collection of lists, it provides an all-encompassing overview of the many aspects and factors relevant to the successful implementation of digital mental health support systems. It hardly leaves anything unmentioned, making the thesis a strong representation of the current state of the field.

In conclusion, this thesis has challenged the existing assumption that AI cannot have a place in mental health care. By exploring user needs and experiences, we identified the necessary requirements for effectively implementing AI-based mental health support systems. These requirements led us to specific implementation needs. After assessing their feasibility, we proposed a practical work plan using AI to implement those needs. Throughout this research, we have suggested methods for AI to aid students in managing and improving their mental health and provided an ethical and patient-forward approach to create an effective implementation. Moving forward, we aim to inspire other researchers to further develop technological solutions from a patient-forward perspective and help change Ana's life for the better.

References

Abd-Alrazaq, A. A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., & Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: Scoping review. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 23(1), e17828. <https://doi.org/10.2196/17828>

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>

Apolinário-Hagen, J., Vehreschild, V., & Alkoudmani, R. M. (2017). Current Views and Perspectives on E-Mental Health: an exploratory survey study for understanding public attitudes toward Internet-Based psychotherapy in Germany. *JMIR Mental Health*, 4(1), e8. <https://doi.org/10.2196/mental.6375>

Auerbach, R. P., Alonso, J., Axinn, W. G., Cuijpers, P., Ebert, D. D., Green, J. G., Hwang, I., Kessler, R. C., Liu, H., Mortier, P., Nock, M. K., Pinder-Amaker, S., Sampson, N. A., Aguilar-Gaxiola, S., Al-Hamzawi, A., Andrade, L. H., Benjet, C., Caldas-De-Almeida, J. M.,

Demyttenaere, K., . . . Bruffaerts, R. (2016). Mental disorders among college students in the World Health Organization World Mental Health Surveys. *Psychological Medicine*, 46(14), 2955–2970. <https://doi.org/10.1017/s0033291716001665>

Borghouts, J., Eikey, E. V., Mark, G., De Leon, C., Schueller, S. M., Schneider, M., Stadnick, N., Zheng, K., Mukamel, D. B., & Sorkin, D. H. (2021). Understanding mental health app use among community college Students: Web-Based Survey study. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 23(9), e27745. <https://doi.org/10.2196/27745>

Borghouts, J., Eikey, E., Mark, G., De Leon, C., Schueller, S. M., Schneider, M., Stadnick, N., Zheng, K., Mukamel, D., & Sorkin, D. H. (2021a). Barriers to and facilitators of user engagement with digital Mental Health Interventions: Systematic review. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 23(3), e24387. <https://doi.org/10.2196/24387>

D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., Gleeson, J., & Alvarez-Jimenez, M. (2017). Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.00796>

Depression treatments for adults. (n.d.). <https://www.apa.org>. <https://www.apa.org/depression-guideline/adults>

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (WoeBot): a randomized controlled trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>

Garrido, S., Millington, C., Cheers, D., Boydell, K., Schubert, E., Meade, T., & Nguyen, Q. V. (2019). What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. *Frontiers in Psychiatry*, 10. <https://doi.org/10.3389/fpsyg.2019.00759>

Graham, A. K., Lattie, E. G., Powell, B. J., Lyon, A. R., Smith, J. D., Schueller, S. M., Stadnick, N. A., Brown, C. H., & Mohr, D. C. (2020). Implementation strategies for digital mental

health interventions in health care settings. *American Psychologist*/the *American Psychologist*, 75(8), 1080–1092. <https://doi.org/10.1037/amp0000686>

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Current Psychiatry Reports/Current Psychiatry Reports*, 21(11). <https://doi.org/10.1007/s11920-019-1094-0>

Grist, R., Croker, A., Denne, M., & Stallard, P. (2018). Technology Delivered Interventions for Depression and Anxiety in children and Adolescents: A Systematic review and meta-analysis. *Clinical Child and Family Psychology Review*, 22(2), 147–171. <https://doi.org/10.1007/s10567-018-0271-8>

Grossard, C., Grynspan, O., Serret, S., Jouen, A., Bailly, K., & Cohen, D. (2017). Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (ASD). *Computers and Education/Computers & Education*, 113, 195–211. <https://doi.org/10.1016/j.compedu.2017.05.002>

Holzmeier, C. (2021). Beyond ‘AI for Social Good’ (AI4SG): social transformations—not tech-fixes—for health equity. *Interdisciplinary Science Reviews/ISR. Interdisciplinary Science Reviews*, 46(1–2), 94–125. <https://doi.org/10.1080/03080188.2020.1840221>

Hudson, G., Negbenose, E., Neary, M., Jansli, S. M., Schueller, S. M., Wykes, T., & Jilka, S. (2022). Comparing professional and consumer ratings of Mental Health apps: Mixed Methods study. *JMIR Formative Research*, 6(9), e39813. <https://doi.org/10.2196/39813>

Ibrahim, I., Mansor, N. H., & Bidin, J. (2022). Factors affecting mental illness and social stress in students using Fuzzy TOPSIS. *Journal of Computing Research and Innovation*, 7(2), 88–100. <https://doi.org/10.24191/jcrinn.v7i2.294>

Inal, Y., Wake, J. D., Guribye, F., & Nordgreen, T. (2020). Usability Evaluations of Mobile Mental Health Technologies: Systematic Review. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 22(1), e15337. <https://doi.org/10.2196/15337>

Javaid, S. (2024, February 16). *Guide to RLHF in 2024*. AIMultiple: High Tech Use Cases & Tools to Grow Your Business. <https://research.aimultiple.com/rlhf/>

Josifoski, M., Sakota, M., Peyrard, M., & West, R. (2023). Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2303.04132>

Kurakin, A., Ponomareva, N., Syed, U., MacDermed, L., & Terzis, A. (2023). Harnessing large-language models to generate private synthetic text. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2306.01684>

Lattie, E. G., Adkins, E. C., Winkquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital Mental Health Interventions for Depression, Anxiety, and Enhancement of Psychological Well-Being among College Students: Systematic Review. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 21(7), e12869. <https://doi.org/10.2196/12869>

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial intelligence for Mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, 6(9), 856–864. <https://doi.org/10.1016/j.bpsc.2021.02.001>

Lovejoy, C. A. (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>

Mohr, D. C., Weingardt, K. R., Reddy, M., & Schueller, S. M. (2017). Three Problems With Current Digital Mental Health Research . . . and Three Things We Can Do About Them. *Psychiatric Services*, 68(5), 427–429. <https://doi.org/10.1176/appi.ps.201600541>

Naslund, J. A., Aschbrenner, K. A., Kim, S. J., McHugo, G. J., Unützer, J., Bartels, S. J., & Marsch, L. A. (2017). Health behavior models for informing digital technology interventions for individuals with mental illness. *Psychiatric Rehabilitation Journal*, 40(3), 325–335. <https://doi.org/10.1037/prj0000246>

Organisciak, P., & Ryan, M. (2022). Improving text relationship modelling with artificial data. *Journal of Information Science*, 50(2), 434–446. <https://doi.org/10.1177/01655515221093031>

Oyebode, O., Alqahtani, F., & Orji, R. (2020). Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. *IEEE Access*, 8, 111141–111158. <https://doi.org/10.1109/access.2020.3002176>

Pedrelli, P., Nyer, M., Yeung, A., Zulauf, C., & Wilens, T. (2014). College Students: Mental health problems and treatment Considerations. *Academic Psychiatry*, 39(5), 503–511. <https://doi.org/10.1007/s40596-014-0205-9>

Place, S., Blanch-Hartigan, D., Rubin, C., Gorrostieta, C., Mead, C., Kane, J., Marx, B. P., Feast, J., Deckersbach, T., Pentland, A., Nierenberg, A., & Azarbayejani, A. (2017). Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 19(3), e75. <https://doi.org/10.2196/jmir.6678>

Pokhrel, P., Karmacharya, R., Salisbury, T. T., Carswell, K., Kohrt, B. A., Jordans, M. J. D., Lempp, H., Thornicroft, G., & Luitel, N. P. (2021). Perception of healthcare workers on mobile app-based clinical guideline for the detection and treatment of mental health problems in primary care: a qualitative study in Nepal. *BMC Medical Informatics and Decision Making*, 21(1). <https://doi.org/10.1186/s12911-021-01386-0>

Prakash, A. V., & Das, S. (2020). Intelligent Conversational Agents in Mental Healthcare Services: A thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, 12, 1–34. <https://doi.org/10.17705/1thci.12201>

Radovic, A., Kirk-Johnson, A., Coren, M., George-Milford, B., & Kolko, D. (2022). Stakeholder perspectives on digital behavioral health applications targeting adolescent depression and suicidality: Policymaker, provider, and community insights. *Implementation Research and Practice*, 3, 263348952211207. <https://doi.org/10.1177/26334895221120796>

Ramu, K., Ramachandran, M., & Sivaji, C. (2023). Understanding Long Short-Term Memory LSTM Models in IBM SPSS Statistics. *Journal on Innovations in Teaching and Learning*, 2(1). <https://doi.org/10.46632/jitl/2/1/3>

Ryan, M., Antoniou, J., Brooks, L., Jiya, T., Macnish, K., & Stahl, B. (2019). Technofixing the future: Ethical side effects of using AI and big data to meet the SDGs. *2019 IEEE SmartWorld*,

Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI). <https://doi.org/10.1109/smartworld-uic-atc-scalcom-iop-sci.2019.00101>

Sholevar, F., Butryn, T., Bryant, L., & Marchionni, C. (2017). The shortage of psychiatrists and other mental health providers: Causes, current state, and potential solutions. *International Journal of Academic Medicine*, 3(1), 5. https://doi.org/10.4103/ijam.ijam_49_17

Sano, A., Taylor, S., McHill, A. W., Phillips, A. J., Barger, L. K., Klerman, E., & Picard, R. (2018). Identifying objective physiological markers and Modifiable behaviors for Self-Reported Stress and Mental Health Status using wearable sensors and mobile phones: observational study. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, 20(6), e210. <https://doi.org/10.2196/jmir.9410>

Schueller, S. M., Hunter, J. F., Figueroa, C., & Aguilera, A. (2019). Use of digital mental health for marginalized and underserved populations. *Current Treatment Options in Psychiatry*, 6(3), 243–255. <https://doi.org/10.1007/s40501-019-00181-z>

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700. <https://doi.org/10.1016/j.ijinfomgt.2023.102700>

Tal, A., & Torous, J. (2017). The digital mental health revolution: Opportunities and risks. *Psychiatric Rehabilitation Journal*, 40(3), 263–265. <https://doi.org/10.1037/prj0000285>

Titov, N., Hadjistavropoulos, H. D., Nielssen, O., Mohr, D. C., Andersson, G., & Dear, B. F. (2019). From Research to Practice: Ten Lessons in Delivering Digital Mental Health Services. *Journal of Clinical Medicine*, 8(8), 1239. <https://doi.org/10.3390/jcm8081239>

Vlek, R. J., Steines, D., Szibbo, D., Kübler, A., Schneider, M., Haselager, P., & Nijboer, F. (2012). Ethical Issues in Brain–Computer Interface Research, Development, and Dissemination. *Journal of Neurologic Physical Therapy*, 36(2), 94–99. <https://doi.org/10.1097/npt.0b013e31825064cc>

Vo, V., Auroy, L., & Sarradon-Eck, A. (2019). Patients' perceptions of MHealth Apps: Meta-Ethnographic Review of Qualitative Studies. *JMIR Mhealth and Uhealth*, 7(7), e13817. <https://doi.org/10.2196/13817>

Weiner, S. J., Schwartz, A., Sharma, G., Binns-Calvey, A., Ashley, N., Kelly, B., Dayal, A., Patel, S., Weaver, F. M., & Harris, I. (2013). Patient-Centered Decision making and health care Outcomes. *Annals of Internal Medicine*, 158(8), 573. <https://doi.org/10.7326/0003-4819-158-8-201304160-00001>

Wykes, T., Lipshitz, J., & Schueller, S. M. (2019). Towards the Design of Ethical Standards Related to Digital Mental Health and all Its Applications. *Current Treatment Options in Psychiatry*, 6(3), 232–242. <https://doi.org/10.1007/s40501-019-00180-0>